FOCUS ON STRUCTURAL BIOLOGY

Mass Spectrometry and Genomic Analysis

Edited by
J. Nicholas Housby

Kluwer Academic Publishers

Mass Spectrometry and Genomic Analysis

Edited by

J. NICHOLAS HOUSBY
Oxagen Limited, Abingdon, United Kingdom

KLUWER ACADEMIC PUBLISHERS
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47595-2
Print ISBN: 0-7923-7173-9

Created in the United States of America

Visit Kluwer Online at: http://kluweronline.com
and Kluwer's eBookstore at: http://ebooks.kluweronline.com

TABLE OF CONTENTS

CHAPTER 3

Hubert Köster

MASSARRAY™: Highly Accurate and Versatile High Throughput Analysis of Genetic Variations

CHAPTER 4

S. Sauer, D. Lechner, IG. Gut

The GOOD Assay

CHAPTER 5

CHAPTER 6

CHAPTER 7

WV. Bienvenut, M. Müller, PM. Palagi, E. Gasteiger, M. Heller, E. Jung, M. Giron, R. Gras, S. Gay, PA. Binz, G J. Hughes, JC. Sanchez, RD. Appel, DF. Hochstrasser

Proteomics and Mass Spectrometry: Some Aspects and Recent Developments

INTRODUCTION

The human genome project has created intense interest from academics, commercial business and, not least, the general public. This is not surprising, as understanding the genetic make up of each individual gives us clues as to the genetic factors that predispose one to a particular genetic disease. In this way the human genome sequence is set to revolutionise the way we treat people for genetic diseases and/or predict patients future health regimes. Single Nucleotide Polymorphisms (SNPs), single base changes in the nucleotide DNA sequence of individuals, are thought to be the main cause of genetic variation. It is this variation that is so exciting as it underpins the way(s) in which the human body can respond to drug treatments, natural defence against disease susceptibility or the stratification of the disease in terms of age of onset or severity. These SNPs can be either coding (cSNP), appearing within coding regions of genes or in areas of the genome that do not encode for proteins. The coding cSNPs may alter the amino acid protein sequence which in turn may alter the function of that particular protein. Much effort is directed towards identifying the functions of SNPs, whether that be within genes (cSNPs) or within regulatory regions (eg. promoter region) that affect the level of transcription of the gene into mRNA.

If an SNP is proved to be truly polymorphic, i.e. it appears in many samples of the population, then individuals can be genotyped for the homozygous form of the allele, the same variation on both chromosomes, or a heterozygous form with a different variation of the SNP on each chromosome. An international SNP working group has been set up to map all of the known human SNPs, it is envisaged that every single gene in the human genome will have a variation within or close to it. By comparing patterns of SNP allele frequencies between disease affected and control populations, disease associated SNPs can be identified and potential disease gene(s) located. These types of study require genotyping of thousands of SNPs which requires the use of powerful, high throughput, systems of analysis. There are many competing new technology platforms which attempt this but the one that 'stands out from the crowd' is mass spectrometry. This book contains a collection of descriptions of some of the most outstanding advances in this field of mass spectrometry (chapters 1-6), from which, I hope, the reader will be able to learn both the principles and the most up to date methods for its use.

Analysis of the proteins produced from mRNA will lead to another level of information analysis. Not all of the proteins produced from mRNA correlates to its expression. Many proteins have alterations at the post-translational stage, mostly by glycosylation or phosphorylation events. It is this that may cause alteration in function of the protein product. It is therefore necessary to investigate at both the gene level and at the protein level. The study of proteomics, the comprehensive study of proteins in a given cell, is discussed in chapter 7. This gives the reader a broader perspective in the uses of mass spectrometry in this fast changing analytical environment of genome research.

J. NICHOLAS HOUSBY

PREFACE

My interest in mass spectrometry stemmed from working in the laboratory of Professor Edwin Southern at the department of Biochemistry, Oxford University, UK. It was there that I was given an ambitious project which involved the analysis of arrays of nucleic acids using mass spectrometry. I must certainly thank him for his tremendous insights into this field and for stimulating my interest in this area of research. Having now moved on from Professor Southern's lab I have become extremely interested in the use of novel technologies for genetic analysis. I am convinced, that over the next decade, mass spectrometry will lead the way in polymorphism screening, genotyping and in other genetic testing environments. It is for this reason that I have put together this book. I have attempted to bring together descriptions, from some of the world leaders in this field of research, of the most recent advances in genomic analysis using mass spectrometry. I make no attempt to make this an exhaustive collection but a text that will 'whet' the appetite of those interested in this fast moving and provocative arena. The final chapter describes the use of mass spectrometry in proteomics, the comprehensive (high throughput) study of proteins in cells. I think that this is a necessary addition for the reader to have a broader insight into the current uses of mass spectrometry in research and development. I hope that this book will be a useful companion to investigators already at the 'cutting edge' but also a guide to those who are interested in learning more about this powerful analytical tool.

J. NICHOLAS HOUSBY

CHAPTER 1

SINGLE-NUCLEOTIDE POLYMORPHISM ANALYSIS BY MALDI-TOF MASS SPECTROMETRY

1. Analysis of Peptide Nucleic Acid Hybridisation Probes

2. Direct Analysis of Invasive Cleavage Products

T.J. Griffin and L.M. Smith
Department of Chemistry, University of Wisconsin-Madison, 1101 University Avenue, Madison, WI 53706-1396. Tel:608-263-2594; Fax:608-265-6780; E-mail smith@chem.wisc.edu

1. INTRODUCTION

As the sequencing of the human genome draws near to completion, it has become evident that there is substantial variation in DNA sequence between any two individuals at many points throughout the genome. Sequence variation most commonly occurs at discrete, single-nucleotide positions referred to as single-nucleotide polymorphisms (SNPs), which are estimated to occur at a frequency of approximately one per 1000 nucleotides [1-4]. SNPs are biallelic polymorphisms, meaning that the nucleotide identity at these polymorphic positions is always constrained to one of two possibilities in humans, rather than the four nucleotide possibilities that could occur in principle [4].

SNPs are important to genetic studies for several reasons: First, a subset of SNPs occur within protein coding sequences [3, 4]. The presence of a specific SNP allele may be implicated as a causative factor in human genetic disorders, so that screening for such an allele in an individual may allow the detection of a genetic predisposition to disease. Second, SNPs can be used as genetic markers for use in genetic mapping studies [2-5], which locate and identify genes of functional importance. It has been proposed that a set of 3,000 biallelic SNP markers would be sufficient for whole-genome mapping studies in humans; a map of 100,000 or more SNPs has been proposed as an ultimate goal to enable effective genetic mapping studies in large populations [6]. Therefore, technologies capable of genotyping thousands of SNP markers from large numbers of individual DNA samples in an accurate, rapid and cost-effective manner are needed to make these studies feasible.

J.N. Housby (ed.), Mass Spectrometry and Genomic Analysis, 1-15.

1.1. MALDI-TOF MS

Among the more promising technologies for SNP genotyping is matrix-assisted laser desorption/ionisation (MALDI) time-of-flight (TOF) mass spectrometry (MS) [7]. Introduced in 1988 by Karas and Hillenkamp [8], MALDI revolutionised the mass analysis of large biomolecules. MALDI-TOF MS has several advantages for analysing nucleic acids, including speed, in that ionisation, separation by size, and detection of nucleic acids takes milliseconds to complete. As signals from multiple laser pulses (~20-100 pulses) are usually averaged to obtain a final mass spectrum, the total analysis time can take as little as 10 seconds. By contrast, conventional electrophoretic methods for separating and detecting nucleic acids can take hours to complete. Additionally, the results are absolute, being based on the intrinsic property of mass-to-charge ratio (m/z). This is inherently more accurate than electrophoresis-based or hybridisation-array-based methods, which are both susceptible to complications from secondary structure formation in nucleic acids. Furthermore, the absolute nature of detection, combined with the detection of predominantly single-charged molecular ions, makes the analysis of complex mixtures possible by MALDI-TOF MS. Finally, the complete automation of all steps, from sample preparation through to the acquisition and processing of the data, is feasible [9], giving MALDI-TOF MS great potential for high-throughput nucleic acid analysis applications.

We describe two approaches to SNP analysis by MALDI-TOF MS, one involving the analysis of peptide nucleic acid hybridisation probes [10], and the other the analysis of products of a novel, enzymatic invasive cleavage assay [11].

2. ANALYSIS OF PEPTIDE NUCLEIC ACID HYBRIDISATION PROBES

Peptide nucleic acid (PNA) [12, 13] is a DNA analogue containing the four nucleobases of DNA attached to a neutrally charged amide backbone (Figure 1a) that retains the ability to base-pair specifically with complementary DNA. The neutral backbone confers unique characteristics on the hybridisation of PNA with DNA, including increased thermal stability of the resulting duplex, the ability to hybridise under very low ionic strength conditions and an increased hybridisation specificity for complementary DNA sequences [12-14], making PNA oligomers useful as allele-specific hybridisation probes. PNA is easily analysed by MALDI-TOF MS [15], because the peptide backbone does not fragment, unlike DNA molecules, which may undergo substantial fragmentation during the MALDI process [16]; also, PNA oligomers do not tend to form adducts with metal cations, which is detrimental to MALDI-TOF mass spectrometric analysis [17], because annealing of these oligomers can be done in buffers containing low salt concentrations and also the neutral amide backbone does not have the tendency to bind to cations that may be present to the same extent as the negatively-charged backbone of DNA.

The approach using PNA hybridisation probes for MALDI-TOF mass spectrometric analysis is comprised of the following steps (Figure 1b): immobilisation of biotinylated target DNA (e.g. a PCR amplicon) by binding to streptavidin coated magnetic beads; dissociation and removal of the non-biotinylated

strand; hybridisation of the PNA probes; washing to achieve proper discrimination; and finally direct analysis by MALDI-TOF MS. During the MALDI process, the PNA probes hybridised to the immobilised DNA targets are dissociated and desorbed from the immobilised target strand, enabling their detection by MALDI-TOF MS, whereas the target DNA remains immobilised on the MALDI probe tip and thus is not detected in the resulting mass spectrum.

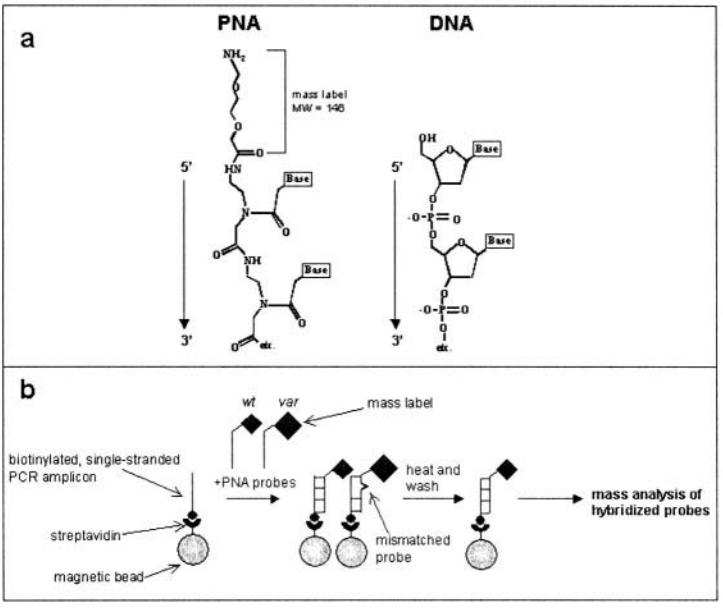


Figure 1. a. Chemical Structures of PNA and DNA. b. MALDI-TOF MS analysis of PNA hybridisation probes.

2.1. Design of PNA Hybridisation Probes

The model system employed in this study was the 182 base pair exon 4 of the human tyrosinase gene. Tyrosinase is a copper-containing enzyme in the melanin biosynthetic pathway. Mutations in the tyrosinase gene have been implicated in type I oculocutaneous albinism. For each of these four polymorphic positions, two allele-specific PNA probes were designed, one complementary to the wild-type allele, the other complementary to the single-base substituted variant allele. Each pair of PNA probes were designated as either *wt* (wild-type sequence) or *var* (variant sequence) along with the corresponding number of the codon in tyrosinase exon 4 where the polymorphic base occurs within each probe sequence. Table 1 shows the sequences

and design of the PNA probes employed in this study. Each probe was uniquely mass labelled to give a distinct, easily resolved, single-charged molecular ion peak when analysed by MALDI-TOF MS. The mass labels attached to the amino terminus of the probes were 8-amino-3,6-dioxaoctanoic acid molecules, each with a molecular weight of 146 Daltons (Figure 1a).

Table 1. PNA probe design. The number of mass labels (X) added to each probe is shown along with the sequence and the mass of the protonated molecular ion. The wt422 probe also contained a fluorescein label. The polymorphic base is shown in bold in each probe.

Probe name	*Sequence*	*M+H*
*wt*419	X_1-TT**G**-GAC-ATA	2621
*var*419	X_2-TT**A**-GAC-ATA	2751
*wt*422	F-X_1-ACC-**G**GG-AAT	2991
*var*422	ACC-**A**GG-AAT	2471
*wt*446	X_4-AGA-TCT-G**G**G	3098
*var*446	X_6-AGA-TCT-G**A**G	3371
*wt*448	X_3-CT-AT**G**-ACT-A	2871
*var*448	X_5-GCT-AT**A**-ACT	3161

2.2. Analysis of Polymorphisms in Tyrosinase Exon 4

For all samples analysed, hybridisation and wash steps with an added pair of PNA probes (wild-type and variant) were performed separately for each of the four polymorphic positions, as was the subsequent MALDI-TOF MS analysis. The separate spectra obtained for each of the four polymorphic positions were then added together to give a final, composite mass spectrum for each sample. In order to initially optimise the hybridisation and wash conditions, control experiments were done using synthetic oligonucleotide targets containing sequences corresponding to the possible alleles at each of the four point mutation positions.

Figure 2 shows representative results obtained from PCR amplicons obtained from two different human genomic DNA samples. Individual 1 was heterozygous at codon 446, and homozygous wild-type for the other three polymorphic positions examined; individual 2 was heterozygous at codon 448 and wild-type at all other positions. These results demonstrate the ability of this approach not only to analyse multiple polymorphic positions on human DNA samples, but also to unambiguously identify heterozygotes, which is critical to effective genetic analysis.

The benefits offered by the use of PNA hybridisation probes in this approach are quite substantial. Not only do they offer a high degree of sequence specificity as described above, but also the ability to hybridise in a buffer containing no salt, which decreases the potential for secondary structure to form in the immobilised DNA target. Additionally, the elimination of salt from the PNA containing solution as well as the decreased tendency of the neutral charged PNA backbone to form salt adducts eliminates the need for extra washing steps which are required to remove salts in DNA based analyses[17]. The results show that the PNA probes give robust,

well-resolved, molecular ion signals in the MALDI-TOF MS analysis, with no base loss, backbone fragmentation or loss of mass labels.

A limitation of this approach lies in the fact that each set of PNA probes requires different wash conditions in order to obtain good discrimination between the wild-

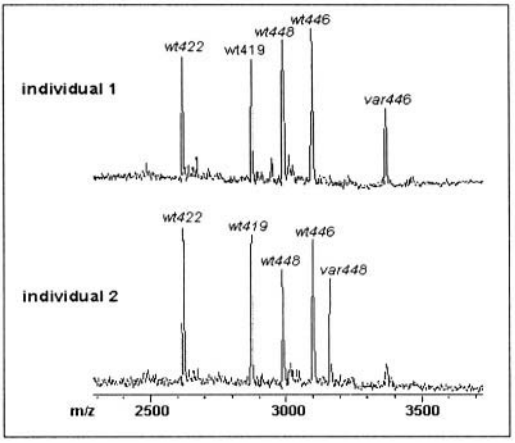


Figure 2. Analysis of human tyrosinase exon 4 polymorphisms in two individuals.

type and variant probes. This is due to highly variable, sequence dependent, thermal stabilities of the duplexes formed between the PNA probes and DNA targets [10]. Optimally, the hybridisation and washing steps for all the polymorphic positions being analysed in an individual sample would be done in one reaction tube, and multiplex MALDI-TOF MS analysis could then be done on one spot on the probe tip. This simultaneous detection of the probes from all of the polymorphic positions would eliminate the need for separate spectra to be taken and then summed together to give a composite spectrum. To this end, approaches to predicting the thermal stabilities of PNA:DNA duplexes have been developed [18, 19] that may allow for the design of PNA probes having similar duplex stabilities, allowing for true multiplex analyses.

3. DIRECT ANALYSIS OF INVASIVE CLEAVAGE PRODUCTS

Common to almost all existing methods of SNP analysis, including the approach described above, is an initial target amplification step using the polymerase chain reaction (PCR), followed by further hybridisation or enzymatic manipulation of the resulting PCR amplicon [2-4, 7]. Despite its widespread utility in basic research, PCR does have significant limitations when used in a high-throughput setting. The fundamental reason for this is the extraordinary sensitivity conferred by the

exponential nature of the PCR process. Although this extreme sensitivity is advantageous for certain applications, it also means that a sample containing no true molecules of a specific sequence that is contaminated by only a few copies of that sequence from another source will amplify the sequence and give a false positive result. As contamination can result from aerosols produced from simply opening a tube or pipetting, laboratories performing high-throughput PCR-based analyses have had to go to extreme lengths to avoid these cross-over contamination problems [20, 21]. Additional issues with the use of PCR for high-throughput analyses include the need for optimisation of each primer set and the corresponding reaction conditions, variability of these reaction conditions between different amplification targets, variability in yields of amplicons produced in different PCR reactions, as well as differential amplification yields of alleles in regions containing sequence polymorphisms [21-23]. Given these inherent limitations to PCR-based high-throughput SNP analysis methods, it is clear that the development of simpler and more direct analysis approaches would be desirable. We describe an alternative MALDI-TOF MS-based approach to analysing SNPs in human DNA that employs the Invader assay [24], an isothermal, highly sequence-specific, linear signal amplification method for the analysis of DNA which does not require an initial PCR amplification of the target sequence.

3.1. The Invader Assay

The Invader assay [24] involves the hybridisation of two sequence-specific oligonucleotides, one termed the Invader oligonucleotide and the other termed the probe oligonucleotide, to a nucleic acid target of interest (Figure 3a). These two oligonucleotides are designed so that the nucleotide on the 3' end of the Invader oligonucleotide (nucleotide "N" in Figure 3a) invades at least one nucleotide into the downstream duplex formed by the probe oligonucleotide and the target strand, forming a sequence overlap at that position. The Invader assay is based on the ability of the 5' nuclease domains of eubacterial Pol A DNA polymerases and structurally homologous DNA repair proteins called Flap endonucleases (FENs) to specifically recognise and efficiently cleave the unpaired region on the 5' end of the probe oligonucleotide, resulting in a 3' hydroxyl terminating DNA cleavage product. Relative to a flap formed by simple non-complementarity of the 5' end of the probe oligonucleotide to the target, a flap that contains sequence overlap between the Invader and probe oligonucleotide is cleaved at a dramatically enhanced rate 3' of the nucleotide located at the position of overlap [25]. Additionally, while the nucleotide at the position of overlap contained in the probe oligonucleotide has a strict requirement of complementarity to the target, the overlapped nucleotide on the 3' end of the Invader oligonucleotide does not have to be complementary to the target for efficient enzymatic cleavage of the 5' flap [24, 25]. The use of thermostable variants of these FENs permits the reaction to be run near the melting temperature (T_m) of the duplex formed between the probe oligonucleotide and target, such that cleaved and uncleaved probe oligonucleotides will cycle off and on the target strand. Thus, with excess probe oligonucleotide present in solution, when a

probe oligonucleotide is cleaved it is replaced by an uncleaved probe oligonucleotide, which is in turn cleaved and replaced, resulting in a linear accumulation of cleavage product with respect to both time and target strand concentration.

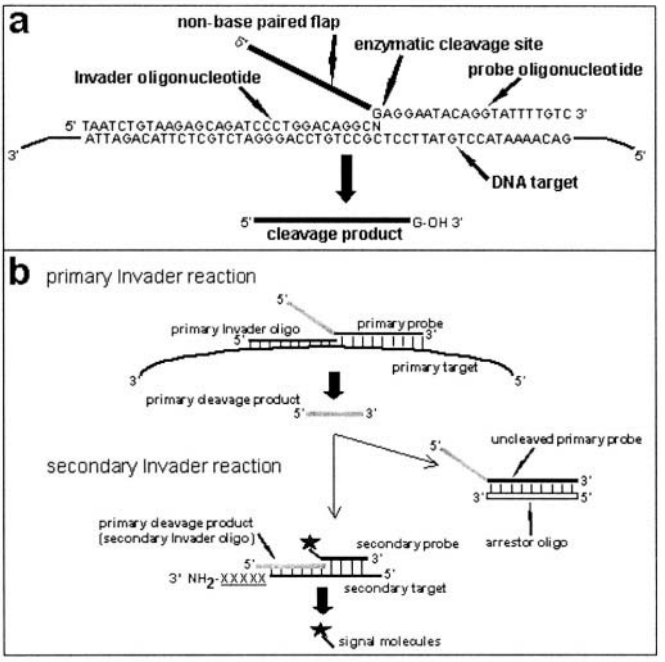


Figure 3. a. Mechanism of the Invader assay. b. The Invader squared reaction.

A modification of the Invader assay, called the Invader squared assay, has also been developed [26] (Figure 3b). The Invader squared assay is a two-step reaction, in which a primary invasive cleavage reaction is directed against a DNA target of interest, producing an oligonucleotide cleavage product as shown in Figure 3a. This cleavage product in turn serves as an Invader oligonucleotide in a secondary invasive cleavage reaction directed against a target oligonucleotide and probe oligonucleotide that are externally introduced into the reaction mix, producing secondary cleavage products (signal molecules) which are then detected. This use of two sequential stages of cleavage reactions approximately squares the amount of amplification of cleavage product compared to a single-step reaction. The Invader squared assay was used in this work to obtain signal at a level necessary for robust detection by MALDI-TOF MS.

Along with the increased amplification of signal molecules when the Invader squared assay is used, there is also an increased potential for the presence of non-specific background signal [26]. One step taken to suppress this background potential was to add an excess of a 2'-O-methyl RNA oligonucleotide to the secondary reaction mix, called the arrestor oligonucleotide [11], that is complementary to the target hybridisation sequence of the primary probe oligonucleotide. This arrestor oligonucleotide anneals to the uncleaved primary probe oligonucleotide molecules present after the primary reaction, rendering the 5' cleavage product sequence, still present on these probe molecules, unavailable to undergo hybridisation with the secondary target. This can lead to background signal accumulation if allowed to occur. 2'-O-methyl RNA nucleotides are not recognised by the FEN, thus ensuring no additional enzymatic cleavage of the structure formed between the arrestor and the probe oligonucleotides. Another step taken to suppress background was to designate the last five nucleotides on the 3' end of the secondary target as 2'-O-methyl RNA (detailed as Xs in Figure 3B), and also to have a 3' amino group, rendering this end of the target inert to the enzyme. This was necessary because the 3' end of the relatively short target has the potential to wrap around and act as the Invader oligonucleotide, displacing the secondary probe oligonucleotide and causing non-specific cleavage and background accumulation of signal molecules.

3.2. Direct Analysis of SNPs From Human Genomic DNA

Figure 4 shows the design of the Invader squared assay employed for the analysis of SNPs in human genomic DNA by MALDI-TOF MS. Figure 4a details the general design of the primary reaction. For any SNP, two allele-specific probe oligonucleotides were designed, each having identical hybridisation sequences complementary to the target DNA. The probe oligonucleotides had different nucleotides at the polymorphic nucleotide position (indicated by the asterisk in the target DNA) which are designated in Figure 4 as "X" and "Y", each being complementary to one of the two possible nucleotides at the SNP position. The nucleotide sequences 5' of X and Y in the probe oligonucleotides were not complementary to the target DNA, and were designed specifically for use in the secondary Invader reaction. The Invader oligonucleotide was designed to be complementary to the target upstream of the probe oligonucleotide region, with a one nucleotide invasion into the probe base-pairing region at the SNP position, so that enzymatic cleavage occurs immediately 3' of nucleotide X or Y in the probe oligonucleotide. This design confers three-fold specificity for SNP detection. First, the Invader oligonucleotide must be complementary to the target and anneal to form the correct overlap structure with the correctly annealed probe oligonuclaotide; second, the endonuclease used in the Invader assay has a strict requirement of absolute complementarity between the target and the nucleotide that occurs at the overlap position in the probe oligonucleotide. Thus, nucleotides X or Y in the probe oligonucleotide must be perfectly complementary to the target at the SNP position in order for the enzyme to recognise the overlap structure and for cleavage to occur;

third, a mismatch at the polymorphic nucleotide between the probe oligonucleotide and the target is thermodynamically destabilising when the reaction is run near the T_m of the duplex. This highly stringent three-fold specificity resulted in the allele-specific accumulation of cleavage products. If the nucleotide complementary to the allele 1 probe was present, then cleavage product 1 accumulated; if the allele 2 nucleotide was present, cleavage product 2 accumulated; in the case of a heterozygote, both cleavage products accumulated over time at similar rates.

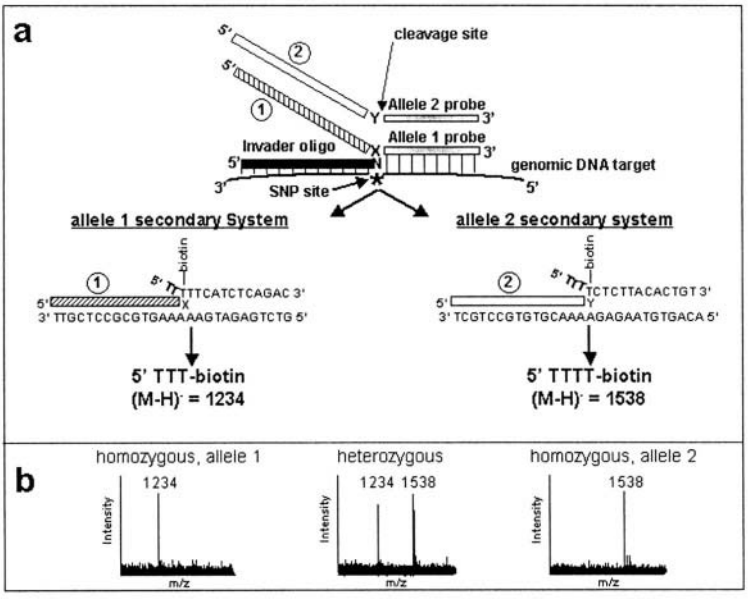


Figure 4. a. Design of the Invader squared assay for MALDI-TOF MS analysis. b. The three possible MALDI-TOF MS outputs for all SNPs analysed by this approach.

After allowing the primary reaction to incubate for two hours, the reaction was decreased and a secondary reaction mix was added that included two allele-specific secondary target oligonucleotides, two secondary probe oligonucleotides, and one arrestor oligonucleotide which annealed to the hybridisation sequence common to each of the primary allele-specific probe oligonucleotides. The sequences of the secondary target and probe oligonucleotides were designed so that the cleavage products from the primary Invader reaction would anneal specifically to one of the secondary targets and act as the Invader oligonucleotide in the secondary reaction. The two allele-specific secondary systems were designed to produce biotinylated signal molecules of unique molecular weights, so that in the subsequent MALDI-TOF MS analysis, the deprotonated, negative, singly-charged molecular ion values detected ($[M-H]^-$ values) would be distinct from each other ($[M-H]^- = 1234$ for allele

1 product and 1538 for allele 2 product). Figure 4b shows the three possible MALDI-TOF MS outputs from this Invader system, corresponding to two possible homozygous genotypes (a single peak at an m/z value of either 1234 or 1538) or a heterozygous genotype (peaks at both m/z values). The same two primary cleavage product sequences shown in Figure 4a were used in every pair of SNP-specific primary probe oligonucleotides, which enabled the use of the same secondary oligonucleotides and signal outputs for each unique SNP analysed. The nucleotides X and Y do not have to be complementary to the secondary target, so primary cleavage products containing any of the four possible nucleotides at the X and Y positions were effective as Invader oligonucleotides in the secondary reaction. A biotin-modified deoxythymidine nucleotide was incorporated in the signal molecules to facilitate solid-phase purification of these molecules using streptavidin coated magnetic beads prior to analysis by MALDI-TOF MS. The signal molecules were designed to contain only deoxythymidine nucleotides because these oligonucleotides are more resistant to fragmentation in the MALDI process than oligonucleotides of other sequences [16].

This approach has proven effective in the analysis of a variety of SNPs in multiple individuals [27]. Figure 5 shows representative MALDI-TOF MS results from the direct analysis of a human genomic DNA sample. All seven of the SNPs analysed by this approach gave unambiguous mass spectral results, showing a single peak in the mass spectrum in the case of homozygous genotypes, and two peaks of approximately equal intensities in the case of heterozygotes. Additionally all different types of SNPs (G to A transitions, G to C transversions, etc.) have been effectively analysed [11, 27].

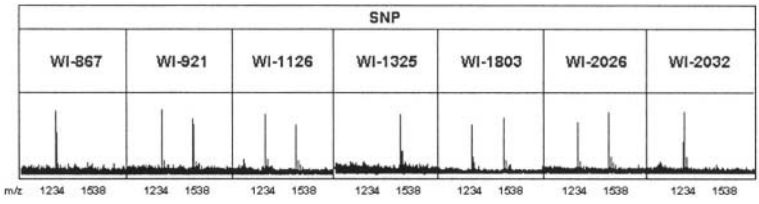


Figure 5. Direct MALDI-TOF MS analysis of seven unique SNPs in a single individual using the Invader assay.

The design of the sequences of the oligonucleotides used in the Invader assay was straightforward, with the only design criteria being that the sequences had thermal duplex stabilities that enabled them to be used at the desired reaction temperature [24, 25]. The primary probe oligonucleotides had hybridisation sequences that were from 16 to 23 nucleotides in length depending on the target sequence, and gave predicted T_ms four to seven degrees above the reaction temperature of 63° C. The primary Invader oligonucleotides were designed to have

a T_m about 15° to 20° C above the corresponding probe oligonucleotides, and were about 30-40 nucleotides in length depending on the target sequence. The secondary reaction oligonucleotides, were designed similarly to work at a reaction temperature of 50° C. As the design of oligonucleotide sequences for use in the Invader assay is simple and robust, the Invader assay should be effective in analysing the vast majority of SNPs found throughout the human genome. As with any method involving oligonucleotide hybridisation, sequences that form significant secondary structures may be problematic, however, because the reaction is run at an elevated temperature some of these problematic sequences may still be effectively analysed.

Integrating the inherent benefits of the Invader assay (highly specific, direct signal amplification without the need for target amplification by PCR) with those conferred by MALDI-TOF MS (extremely rapid and accurate signal detection) represents a significant advance in the development of approaches for the high-throughput genotyping of SNPs. The relatively simple, isothermal Invader assay and the solid-phase sample preparation procedure lend themselves nicely to automated sample handling, giving this approach much potential to the high-throughput genotyping of SNPs for genetic analysis.

4. CONCLUSIONS

We have described two general approaches to SNP analysis by MALDI-TOF MS. Both are designed to incorporate informative signal molecules (PNA hybridisation probes and DNA invasive cleavage products) that are robustly analysed by MALDI-TOF MS and take advantage of the speed and accuracy of this analytical technology. The approach using PNA hybridisation probes is useful for the routine analysis and screening of all types of SNPs from PCR amplicons; the approach involving the Invader assay is ideally suited for the high-throughput analysis of SNPs on a genome-wide scale, useful in a wide variety of genetic studies.

5. EXPERIMENTAL METHODS

5.1. PNA Probe Synthesis and Preparation

PNA probes were synthesised by Perceptive Biosystems, Framingham, MA. These were purified by RP-HPLC and quantified by UV absorbance at 260 nm. The purity and m/z values of the probes were verified by MALDI-TOF MS.

5.2. PCR Amplification of Exon 4 of the Tyrosinase Gene

The primers 5'-GGAATTCTAAAGTTTTGTGTTATCTCA-3' and 5'-TTAATATATGCCTTATTTTA-3', employed for the amplification of human genomic samples, yields a 347 nt fragment from exon 4 and adjacent intronic sequences. Due to the small amount of genomic DNA sample available, these products were re-amplified by nested-PCR using the primer set 5'-biotin-

CTGAATCTTGTAGATAGCTA-3' and 5'-TATTTTTGAGCAGTGGCTCC-3', and the resulting 182 nt products were analysed.

5.3. Hybridisation of PNA Probes to Immobilised Gene Targets

Purified, double-stranded, biotinylated amplicons from a single PCR amplification reaction were combined with 160 μg of streptavidin Dynabeads M-280 (Dynal, Hamburg, Germany), and allowed to bind for 15 minutes at room temperature in 100 μl of binding buffer (10 mM Tris pH 7.0, 1 M NaCl). These were washed once with 100 μl of binding buffer. 100 μl of 0.1 M NaOH was then added to the beads and dissociation of the double-stranded DNA was allowed to occur for 10 minutes. The beads were washed once with 100 μl of 0.1 M NaOH and then three times with 100 μl of hybridisation buffer (10 mM Tris pH 7.0, no NaCl added) to remove the dissociated, non-immobilised DNA strand.

Each immobilised, single-stranded PCR amplicon sample, containing all four of the tyrosinase exon 4 point mutation targets within its sequence, was divided into equal portions in four separate tubes and brought up in 50 μl of hybridisation buffer. One pair of PNA probes was then added to one of the four tubes. The PNA probe pairs were added in the following amounts (pmol WT:pmol VAR): 419-7.5:30; 422-15:7.5; 446-7.5:30; 448-7.5:15. Hybridisation took place for 15 minutes at room temperature. Each reaction tube was then heated for five minutes at the following temperatures, depending on which pair of PNA probes had been added: 419 probes-37 °C; 422 probes-58 °C; 446 probes-58 °C; 448 probes-37 °C. The optimal amounts of each PNA probe added and also the optimal wash temperatures were obtained empirically, in experiments using the immobilised oligonucleotides as targets for the PNA probes. These conditions were considered to be satisfactory if sufficient discrimination between a one-base mismatched target was obtained, as well as approximately equal signal intensity for the two PNA probes when both oligonucleotide targets for a probe pair were present. After this first wash, the supernatant was then removed from each reaction tube, and 50 μl of washing buffer (10 mM Tris, pH 7.0, 0.1% SDS) was then added to the beads and the tubes were heated at their respective temperatures for five minutes, the supernatant removed and the wash repeated for an additional five minutes. The beads were then rinsed once with wash buffer, and once more with ice-cold hybridisation buffer to remove the SDS from the beads. The beads were then brought up in 1 μl of hybridisation buffer. This 1 μl of beads from each reaction tube was then separately spotted on the MALDI probe tip and allowed to dry for approximately 10 minutes. To this, 1.3 μl of matrix (2,5-dihydroxybenzoic acid at 0.02 mg/μl in 9:1 H_2O:acetonitrile) was added and allowed to crystallise. If satisfactory crystals did not form the first time, an additional 0.5 μl of matrix was then added to the beads.

5.4. MALDI-TOF MS Analysis of PNA Probes

Mass spectra were obtained on a Bruker Reflex II time-of-flight mass spectrometer (Billerica, MA), equipped with a 337 nm N_2 laser and operated in the linear,

positive-ion detection mode using delayed extraction with an initial accelerating voltage of 25 kV. For each sample analysed, separate spectra were acquired for each of the four polymorphic positions, and these were then summed together using the mass spectrometer acquisition software to give a composite mass spectrum for each sample. Calibration of the instrument was achieved by use of bovine insulin as an external standard.

5.5. Invader Squared Reaction

All oligonucleotides used were synthesised by the University of Wisconsin Biotechnology Centre (Madison, WI) or Integrated DNA Technologies (Coralville, IA). All probe oligonucleotides used in the primary Invader reaction were PAGE purified. All other oligonucleotides were synthesised with the trityl group on and purified using Sep-Pak C18 reverse-phase purification cartridges (Waters Corp., Milford, MA). Each primary Invader reaction consisted of 3 μL of nuclease-free water, 1 μL of 10X Reaction Buffer (Third Wave Technologies, Madison, WI), 1 μL of 10 μM primary Invader oligonucleotide, and 2 μL of 0.5 μg/μL human genomic DNA in water. This reaction mix was incubated at 95° C for 5 minutes to denature the genomic DNA. The reaction mix was brought to 63° C and immediately 3 μL of a solution containing 75 nanomoles $MgCl_2$, 5 picomoles of each of the two primary probe oligonucleotides, and 100 ng of the *Afu* FEN 1 enzyme (Third Wave Technologies, Madison, WI) was added to give a final reaction volume of 10 μL. This primary reaction was incubated at 63° C for 2 hours. The reaction was then brought to 50°C and the secondary reaction mix (3 μL) was added which contained 40 picomoles of 2'-O-methyl RNA arrestor oligonucleotide, 10 picomoles of each secondary probe oligonucleotide and 0.5 picomoles of each secondary target oligonucleotide. The secondary reaction was incubated at 50° C for 2 hours.

5.6. MALDI-TOF MS Sample Preparation of Cleavage Products

To each completed Invader reaction 100 μg of Dynabeads M-280 streptavidin-coated magnetic beads (Dynal, Oslo, Norway) contained in 120 μL of Immobilisation Buffer (10 mM Tris-HCl, 2 M NaCl, pH 7.0) was added. This solution was mixed well and incubated at room temperature for 10 minutes with gentle shaking. The bead solution was transferred to a 1.5 mL microcentrifuge tube and placed in a Dynal magnetic concentrator (MC). The beads were then washed once with 125 μL of Wash Buffer 1 (10 mM diammonium citrate, 0.1% SDS, pH 7.0) and then twice with 150 μL of Wash Buffer 2 (200 mM diammonium citrate). The beads were then resuspended in 150 μL ultra pure deionised water, transferred to a clean 1.5 mL microcentrifuge tube and washed 3 times with 150 μL of ultra pure water. The washed beads were then resuspended in 40 μL of freshly prepared Elution Buffer (1:1 CH_3OH:NH_4OH[30%]) and incubated at 60° C for 10 minutes. After this incubation, the microcentrifuge tube was immediately placed in the MC and the supernatant was removed and transferred to a clean tube, being careful to remove the magnetic beads as completely as possible. The volatile Elution Buffer

was then completely removed by centrifugation under vacuum for about 15 minutes. The clean, dry sample was then resuspended in 1 μL of 1:1 CH_3CN:ultrapure water.

5.7. MALDI-TOF MS Analysis of Cleavage Products

1 μL of MALDI matrix (1% α-cyano-4-hydroxycinnaminic acid in 1:1 CH_3CN:ultrapure water) was spotted on the MALDI sample plate and allowed to air-dry. To the dried matrix crystals, the resuspended sample in 1μL of 1:1 CH_3CN:ultrapure water was added and allowed to air dry. MALDI-TOF MS analysis was done on a Perceptive Biosystems (Framingham, MA) Voyagir DE-STR mass spectrometer using a nitrogen laser at 337 nm with an initial accelerating voltage of 20 kV and a delay time of 100 nanoseconds. The instrument was! run in reflector mode using negative ion detection with external instrument calibration. All spectra acquired consisted of averaged signal from 50-100 laser shots and the data was processed using accompanying Perceptive Biosystems mass spectrometry software.

6. AFFILIATIONS

All work was conducted at the Department of Chemistry, University of Wisconsin, Madison, 1101 University Avenue, Madison, WI 53706.

7. REFERENCES

1. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lander ES, et al. *Science* 280: 1077, 1998
2. Schafer AJ, Hawkins JR. *Nature Biotechnol* 16: 33, 1998
3. Landegren U, Nilsson M, Kwok PY. *Genome Res* 8: 769, 1998.
4. Brookes AJ. *Gene* 234: 177, 1999.
5. Kruglyak L. *Nature Genet* 17: 21, 1998.
6. Collins FS, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L. *Science* 282: 682, 1998
7. Griffin TJ, Smith LM. *Trends Biotechnol* 18: 77, 2000
8. Karas M, Hillenkamp F. *Anal Chem* 60: 2299, 1988
9. Van Ausdall DA, Marshall WS. *Anal Biochem* 256: 220, 1998
10. Griffin TJ, Tang W, Smith LM. *Nat Biotechnol* 15: 1368, 1997
11. Griffin TJ, Hall JG, Prudent JR, Smith LM. *Proc Natl Acad Sci USA* 96: 6301, 1999
12. Egholm M, Buchardt O, Christensen L, Behrens C, Freier SM, Driver DA, Berg RH, Kim SK, Norden B, Nielsen PE. *Nature 365*: 566, 1993
13. Corey DR. *Trends Biotechnol* 15: 224, 1997
14. Tomac S, Sarkar M, Ratilainen T, Wittung P, Nielsen P, Norden B, Graslund A. *J Am Chem Soc*118: 5544, 1996
15. Butler J, Jiang-Baucom P, Huang M, Belgrader P, Girard J. *Anal Chem* 68: 3283, 1996
16. Zhu L, Parr G, Fitzgerald M, Nelson C, Smith LM. *J Am Chem Soc* 117: 6048, 1995
17. Shaler TA, Wickham JN, Sannes KA, Wu KJ, Becker CH. *Anal Chem* 68: 576, 1996
18. Griffin TJ, Smith LM. *Anal Biochem* 260: 56, 1998
19. Giesen U, Kleider W, Berding C, Geiger A, Orum H, Nielsen PE. *Nucleic Acids Res* 26: 5004, 1998
20. Erlich GD. in *PCR-based Diagnostics in Infectious Disease,* Blackwell Scientific Publications, pp. 3-18, 1994
21. Kwok S, Higuchi R. *Nature* 339: 237, 1989

22. Farrell RE. *Immunol Invest* 26: 3, 1997
23. Vaneechoutte M, Van Eldere J. *J Med Microbiol* 46: 188, 1997
24. Lyamichev V, Mast AL, Hall JG, Prudent JR, Kaiser MW, Takova T, Kwiatkowski RW, Sander TJ, de Arruda M, Arco DA, Neri BP, Brow MA. *Nature Biotechnol* 17: 292, 1999
25. Lyamichev V, Brow MA, Varvel VE, Dahlberg JE. *Proc Natl Acad Sci U S A* 96: 6143, 1999
26. Hall JG, Eis PS, Law SM, Reynaldo LP, Prudent JR, Marshall DJ, Allawi HT, Mast AL, Dahlberg JE, Kwiatkowski RW, de Arruda M, Neri BP, Lyamichev VI. *Proc Natl Acad Sci USA* 97: 8272, 2000
27. Griffin TJ, Smith LM. *Anal Chem* 72: 3298, 2000

CHAPTER 2

SNP GENOTYPING BY MALDI-TOF MASS SPECTROMETRY

LA. Haff, AC. Belden, LR. Hall, PL. Ross, IP. Smirnov
Applied Biosystems, 500 Old Connecticut Path, Framingham MA01701 USA. Tel:508-383-7459; Fax:508-383-7883; E-mail: haffla@appliedbiosystems.com

1. INTRODUCTION

Single-nucleotide polymorphisms, or SNPs, are defined as single base changes in the genome that occur at a frequency greater than 1% in the general population [1]. There is currently great interest in SNPs, partly because they are highly useful in linkage mapping and partly because some SNP variants may contribute significantly to common diseases. SNPs occur in both coding and non-coding regions at intervals of about 1 in 350 bases, with an average frequency of the minor allele at an SNP site of 3% [1]. SNPs are generally biallelic, although in a recent study 3 of 267 SNPs were found to be homozygous for a third allele [1]. SNPs are being discovered at a rapid pace: the SNP Consortium has identified over 100,000 SNPs and Celera Genomics has recently launched an SNP database with 2.8 million unique SNPs. Although only about one tenth of these SNPs lie within coding regions, it is likely that some SNPs in non-coding regions lie within sequences involved with regulatory functions.

The importance of SNPs lies in their direct or indirect association with phenotypes such as susceptibility to disease or an individual's response to a drug. A linkage between an SNP and a phenotype is generally supported through association studies in which a population of known phenotypes is genotyped to determine if the frequency of appearance of an allele is increased in affected individuals. For example, the APO E4 genotype has been closely associated with Alzheimer's disease [2] and likewise the Factor V Leiden allele has been associated with deep-venous thrombosis [3]. The relatively high density of SNPs throughout the genome, but the relatively low frequency of the minor alleles, requires that a substantial population of individuals must be tested against relatively large numbers of putative SNPs to establish an association of an SNP and a phenotype.

Many putative SNPs are discovered as a heterozygous locus from a single individual, as indicated by conventional DNA sequencing or techniques such as denaturing high performance liquid chromatography [4], single-stranded conformational polymorphism electrophoresis or allele-specific hybridisation. Others are established by comparing the sequence at a given locus between a few individuals. However, the percentage of false positive SNPs discovered by these

J.N. Housby (ed.), Mass Spectrometry and Genomic Analysis, 16-32.

techniques is typically about 40% [1]. A valid SNP can often be distinguished from a false positive by repeated automated DNA sequencing, preferably with a reversed orientation primer. However, neither validation nor routine analysis of SNPs by automated DNA sequencing is particularly fast or economical. The Sequazyme™ Pinpoint SNP Typing assay (hereafter referred to as the SNP Typing Assay) was designed for such SNP validation and correlation studies.

2. SNP ANALYSIS BY SINGLE BASE EXTENSION OF PRIMERS

A wide array of SNP typing assays exists, including kinetic PCR (TaqMan®), allele-specific hybridisation, ligase chain reaction, chemical and enzymatic mismatch cleavage, invasive cleavage assays, and others [5]. Single base primer extension, often termed mini-sequencing, is a core technology that has been married to a number of different means of detection. In single base primer extension assays, a primer sequence is selected in which the 3' base of the genotyping primer terminates one base immediately upstream from the complementary base on the target sequence. A suitable DNA polymerase then adds a single ddNMP to the primer, which is complementary to the target sequence, from a pool of all four ddNTPs (Figure 1). The base added can be determined from a label attached to it, such as a fluorophore. However, one can more directly determine the base added by analysis in a mass spectrometer. In this case, no signal molecule is required because the base is revealed by its inherent mass.

A key advantage of primer extension assays is that single base extensions by DNA polymerases generally have low error rates; typically well below the limit of detection, under a wide range of environmental conditions. In contrast, the accuracy of genotyping methods involving competitive annealing of allele-specific probes is generally more problematic. This is due to the complexities of DNA duplex melting behaviour due to sequence variation and the consequent difficulty of obtaining adequate specificity under a single set of environmental conditions.

Several modes of detecting a single base extension have been investigated, including measuring the incorporation of fluorescent, haptenated, or radioactive ddNTPs [6, 7] or gel electrophoresis based-detection of fluorescent primers extended by non-fluorescent nucleotides [8]. Recently, the Applied Biosystems SNapshot™ Primer Extension Kit was introduced. In this assay, a primer is extended by one or more fluorescent- labelled dideoxynucleotides with subsequent detection in a fluorescent-based DNA sequencer. Several primers can be analysed within one lane of the DNA sequencer.

MALDI-TOF mass spectrometry has recently proven an attractive means to analyse single base extensions. Several thousand samples can be analysed each day, because each analysis requires only a few seconds. Furthermore, within each sample, multiple independent loci can be simultaneously analysed. With this internal multiplexing, tens of thousands of SNP genotypes can be obtained each day. Reagent costs are comparatively low, because there are no signal molecules in the assay and the primers are unlabeled and comparatively inexpensive. Because

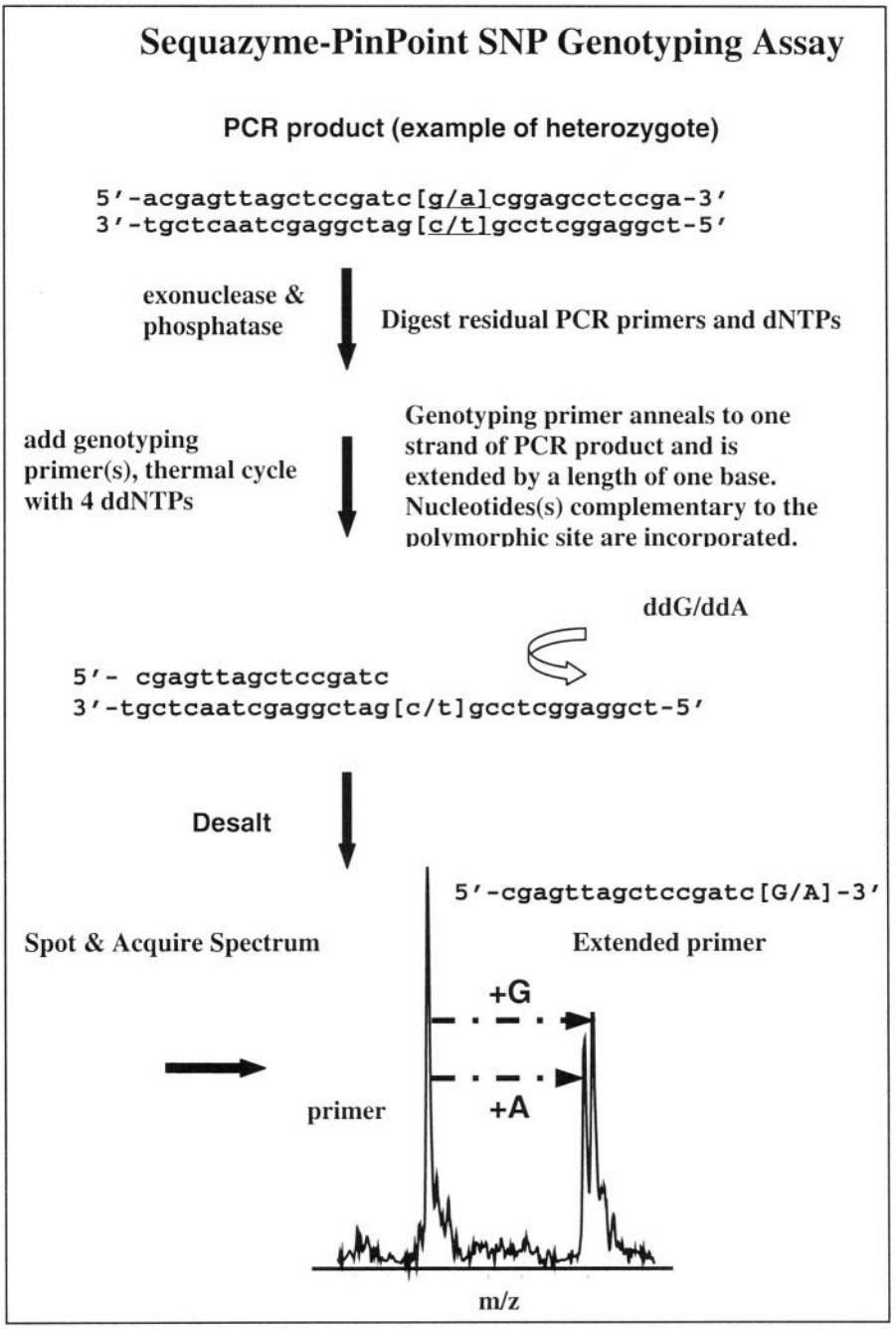


Figure 1. Schematic of the Sequazyme-PinPoint Assay

MALDI-TOF mass spectrometry is very high in resolution, the available degree of multiplexing is much higher than using fluorescent detection, which is typically limited due to a great deal of spectral overlap between different dyes. Allele determinations by mass spectrometry are nearly artefact-free, because the process measures the intrinsic mass of the bases added to the primer. These mass measurements are not affected by secondary considerations, such as the conformation of the DNA, which can confound gel-based techniques.

3. MATERIALS AND METHODS

Materials and methods for conducting the SNP Typing Assay have been previously described in detail [9]. Reagents for the SNP Typing assay, including primer extension, purification and MALDI-TOF reagents, are commercially available in a kit (Sequazyme™ Pinpoint SNP Typing Kit, Applied Biosystems, Foster City, CA). MALDI-TOF mass spectra were acquired with an Applied Biosystems Voyager-DE Biospectrometry™ Workstation, by delayed extraction MALDI-TOF in the positive ion mode.

4. DESIGN CONSIDERATIONS FOR THE SNP GENOTYPING ASSAY

4.1 Design of PCR Product

The normal source of DNA for the SNP Genotyping Assay is polymerase chain reaction (PCR) amplified DNA. Generally, the PCRs can be performed with any one of a variety of thermostable DNA polymerases, including *AmpliTaq*, *AmpliTaq Gold*, Stoffel fragment of *Taq* DNA polymerase, and *rTth* DNA polymerases. Of necessity, the SNP site or sites to be genotyped must lie between the PCR primers. All other factors being equal, superior results are generally obtained with shorter PCR products rather than longer ones. This is because signal intensities in the SNP Genotyping Assay are generally improved with increasing molarity of PCR product, and shorter PCR products are generally produced in higher molar yield. PCR products with only one or two bases between the primers have been found to produce excellent results in the assay, so there is no apparent lower limit to the PCR product size. Because it is possible to produce a single PCR product with multiple SNP sites, it may be tempting to produce large PCR products. However, products over 1,000 bp may be difficult to genotype because they are produced in lower molar yield. It is desirable to produce the PCR product at a concentration of around 5×10^{-8} M; this is a concentration producing a readily discernable band upon gel electrophoretic analysis with ethidium bromide staining. An advantage of the SNP assay is that it does not require a PCR product made with high specificity. The presence of additional, non-specific PCR products in the PCR products generally does not interfere with obtaining a correct genotype, because the genotyping primer provides another level of discrimination against typing non-specific PCR products.

4.2. PCR Product Polishing

PCR polishing is the process of destroying the residual dNTPs and primers, which otherwise would interfere with the primer extension assay (Figure 1). The PCR product itself need not be converted to single stranded form for the SNP assay.

Destruction of the residual PCR primers is not strictly required, because they will not interfere with the genotyping reaction as long as the PCR primers are different in length from all the genotyping primers. Polishing is easily accomplished by adding a mixture of shrimp alkaline phosphatase and exonuclease 1 and incubating in a single thermal cycle reaction of 20 min at 37°C –20 min. 85° C. This step polishes the sample and inactivates the thermolabile enzymes so that they do not interfere with subsequent steps.

4.3. Primer Design Rules for Monoplex SNP Typing

For genotyping with a single primer (monoplex genotyping), the rules for primer design are very simple. The primer sequence must be complementary to one of the PCR product strands, with its 3' hydroxyl end terminating one base upstream (toward 5' end) of the SNP Site (Figure 2). Generally, the primer should be relatively short, to maximise MALDI signal and resolution. The experimentally determined lower limit to primer size is about 15 bases with *Tth* DNA polymerase, and about 12 bases long with *ThermoSequenase* or *AmpliTaq FS* DNA polymerase. Primers need to be of these lengths to extend efficiently on double stranded templates such as PCR products, although primers a few bases shorter will extend efficiently on single stranded templates.

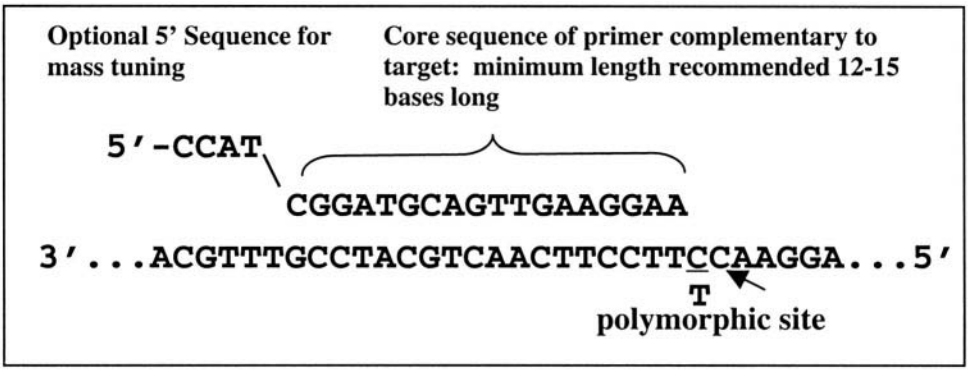


Figure 2. Optimal Design of SNP Genotyping Primers. The primer may be selected to anneal to either strand of the target DNA, with the 3' hydroxyl end of the primer terminating one base upstream of the SNP site. Generally, a minimum core sequence of 15 bases complementary to the target is recommended. Additional 5' bases may be added to the primer sequence to adjust its mass, and these additional bases need not be complementary to the target.

For typical mixed-base primer sequences, base composition and T_m do not appear to be important design criteria for genotyping primers. Virtually all primers extend well by thermal cycling with a denaturation step at 94° C and an extension step at 37° C step [9].

It is good practice to check the genotyping primer sequence with a software program for oligonucleotide sequence analysis to determine that the primer does not contain a hairpin structure. Primers with hairpins of four or more contiguous base pairs may extend poorly, or if the 3' hydroxyl group is recessed, such primers may extend in the absence of target (self-extension). Self-extension should be ruled out by conducting a control extension reaction with the primer in the absence of target. If a primer exhibits self-extension, truncating the sequence may eliminate the hairpin. Otherwise, a primer can be selected for the opposite strand orientation. In our laboratory, between 1-2% of chosen primer sequences have exhibited self-extension, but no sites have been found in which both positive and reverse strand primers self-extended.

Primer sequences containing four or more identical contiguous bases (or repeats) near the 3' end also may exhibit inadequate specificity, because in this case there could be multiple configurations for primer annealing. In such a case it is also desirable to select the alternate polarity primer for the SNP site. Another issue arises if two or more SNP sites are so closely spaced that the genotyping primers partially overlap the same sequence. This is generally not a problem, although there will generally be a slight signal loss in the assay due to the primers competing for the same sequence. However, overlapping primers from complementary strands should be avoided, because they are be more likely to anneal to each other, causing self-extension.

4.4 Mass Calculations

Unextended primer peaks in the spectrum are identified by their mass, and primer extensions by the additional mass added to the primer. The mass of an ordinary oligonucleotide primer is based entirely upon the number of each type of base in the sequence. The mass of a positively charged oligonucleotide primer, in Daltons, is:

$$\text{Mass} = (\#A \times 313.21) + (\#C \times 289.18) + (\#G \times 329.21) + (\#T \times 304.20) - 61.965 \text{ AMU}$$

Many oligonucleotide venders report the mass of custom oligonucleotides, but these values should be used with caution, since many suppliers use incorrect algorithms. Use the vendor's reported molecular weight only if it can be established that the vendor is using the correct algorithm. Oligonucleotides with modifications pose an additional problem, since the above algorithm is insufficient and the vendor often does not supply the correct mass for the modification. In this case it is often best to simply measure the mass of the primer.

4.5. Primer Design Rules for Multiplexed Reactions

Multiplexed reactions contain PCRs of multiple targets, which require multiple genotyping primers for analysis. The multiplexed targets can be a single PCR, to be genotyped at several sites, or several independent PCR reactions mixed together, or a multiplexed PCR. Multiplex reactions increase throughput, but also reduce the cost per analysis. It costs little more to genotype multiple targets in one tube rather than a single one. The only drawback is that multiplexed PCRs generally produce a somewhat lower PCR yield at each site compared with preparing the PCRs individually.

The guidelines for selecting genotyping primers to be mixed together as a pool for multiplexed assays are basically the same as for monoplex assays. However, an additional guideline holds that for pools of primers, the masses of all the primers and all their possible extension products should not overlap. There are two recommended strategies for multiplexing, depending upon how many primers are to be run together in a single pool.

4.5.1. Multiplexing with Primer Pools of Six or Fewer Primers

A simple design rule is to select primers that differ in length from each other by a length of two bases. This corresponds to a mass difference averaging over 600 Daltons apart. With such two-base spacing, it is extremely unlikely that any two primers, regardless of their base composition, will overlap in mass. An example spectrum of a six-fold multiplex SNP genotyping reaction, with two-base T spacing, is illustrated in Figure 3.

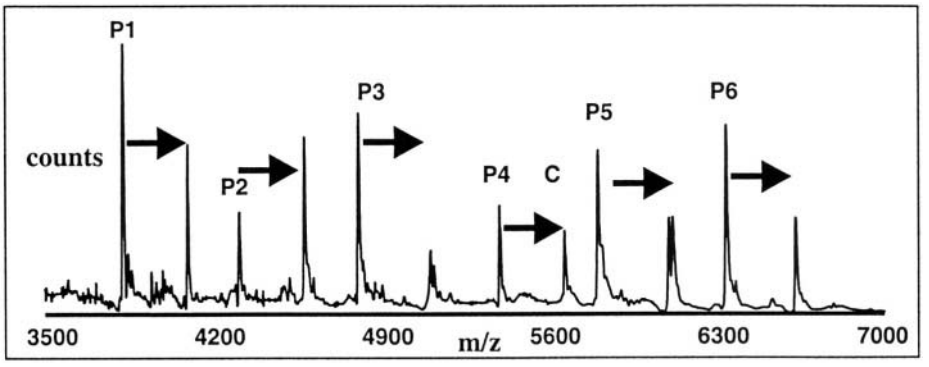


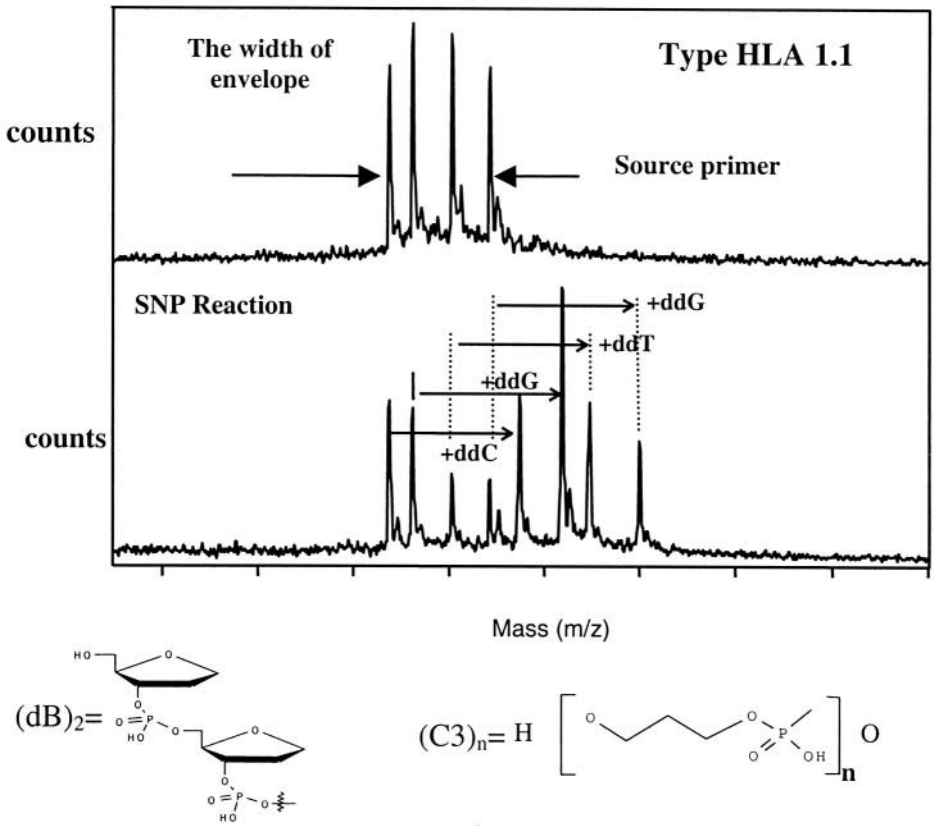
Figure 3. A 6-plex PCR multiplexed by designing primers differing in length by 2 T residues. PCR primers and genotyping primers were designed for the following loci from the Whitehead Institute Centre for Genome Research SNP database: WIAF 445, WIAF 135, WIAF 2083, WIAF 1970, WIAF 1438, and WIAF 1243.

The recommended strategy is to select a 15-base *core primer sequence* for each oligonucleotide primer, that is, a sequence completely complementary to the target sequence. The first primer selected may have the shortest length, typically 15 bases, and when an additional primer is designed for this pool, the length of this additional primer should be incremented by two bases. These two extra bases can be complementary to the target, but need not be, because 5' non-complementary bases 15 bases from the 3' end have little effect upon the stability of the primer-template duplex (figure 2). All other factors being equal, additional T or C bases are recommended. Both these bases are highly resistant to base fragmentation and result in higher quality spectra than if one adds A or G bases, which are subject to depurination. This design rule works well for up to about six primers in a primer pool. If the primer pool contains more than six primers, in which case the longest primer will typically be longer than ~27 bases, reduced sensitivity may be observed from the extensions with the longer primers, because MALDI-TOF signal intensity decreases with increasing primer mass.

4.5.2.Recommended Primer Pool Design: More than six primers

When preparing a primer pool of more than six primers, a superior assay design is to space primers closer together in mass than two bases apart, a design process called *mass tuning*. The idea is to "fit" primers between other primers and their extension products. 1-3 primers can be spaced in mass between a first primer and its extension products. Theoretically, primers can be spaced as close together as 40 Daltons, the difference in mass between the lightest terminator (ddCTP) and the heaviest terminator (ddGTP). In practice, to accommodate peak widths and to minimise overlaps with impurities in the primers, spacing primers 80 –120 Daltons apart works better. Fine-tuning of primer masses can be accomplished by introducing propylene glycol and abasic residues. Figure 4 illustrates an example of typing HLA-DQα PCR product with four sets of primers, all of which were 13 bases long. Through the introduction of the synthetic mass modifying groups, enough mass variation was created to space the primers without overlaps, with only 212 Daltons separating the lightest and heaviest primers. As can be seen, there was little fall-off in signal between the largest and smallest primers and their extensions, and yet every peak was clearly resolved.

The example in Figure 4 also illustrates a further extension of mass tuning, how to design primers when there are additional polymorphic sequences in the target that might create potential mismatches between the genotyping primer and its target. HLA sequences are so highly polymorphic that primers for genotyping the selected sites overlapped known sites of possible additional polymorphic variation. This additional genetic variation could interfere with the genotyping primer annealing to its target sequence. We encountered this problem for three of four of these primer sites. To prevent this possible problem, primers were synthesised which matched each of the two



Primer Sequence (Daltons)		2nd primer in pair		mass
P1	=5′ (dB)$_2$AATTTGATGGAGA-3′	NONE		4400
P2C	=5′ (C3)$_3$TGATGGAGACGAG-3′	P2T	=5′ (C3)$_3$CGATGGAGATGAG-3′	4481
P5A	=5′ (C3)$_2$GTGGACCTGGAGA-3′	P5G	=5′ (C3)$_2$ATGGACCTGGGGA-3′	4318
P8C	=5′-(C3)$_2$TGAGACTGCCTGG-3′	P8C	=5′-(C3)$_2$CGAGACTGTCTGG-3′	4269

Figure 4. Genotyping of HLA-DQα PCR Product, Type 1.1. Seven primers, comprising four different masses containing 5'-mass shifting labels $(dB)_2$ and $(dC3)_n$ (where n=2 or 3), were selected to type HLA PCR product, each primer terminating one base upstream from the site to be typed. Primer P1 did not overlap any of these additional polymorphic sites in the target DNA, but primers P2, P5, and P8 did overlap an additional polymorphism. Primers P2, P5, and P8 ere synthesised as pairs of the same sequence, except for two bases indicated by underlining, one at the 5'end and another at the polymorphic position. An additional terminal 5' base was selected and added to each oligonucleotide sequence in the pair, so that each primer in the pair was of the same base composition and mass.

possible polymorphic sequences in each target sequence. An undesirable side effect of this choice is that the primers differed in base composition and therefore masses. To eliminate this mass variation within pairs of primers, an additional 5' base was added to each primer, which cancelled out this mass difference. This way, three of the four primer peaks in figure 4, P2, P5, and P8 actually represented pairs of equal mass primers. Generally, only one of the pair will be extended, depending upon the underlying polymorphism, but which primer of the pairs extends is not important since it is only the site after the primer being genotyped.

For mass tuning of primer pools, either a spreadsheet program or a specialised primer design software package (i.e. Applied Biosystems) is recommended to prepare a histogram visualising the mass windows occupied by all the primers and their extension products. By preparing this calculated spectrum, one can design the primer pool so that no primers or their extension products can overlap. Masses can be manipulated as desired by adding sequences on the 5' end of the core primer sequence, either complementary to the target sequence or not (Figure 2). Software assisting in primer design and base calling software is highly desirable in dealing with multiplexed reactions. For example, the 12-fold multiplexed reaction illustrated in Figure 5 was designed so no primers or extension products could overlap. It is difficult for the eye to identify which peaks correspond to primers and their extension products, but this identification is easy with software specific to the task. Specialised software for sorting out the unextended primers, automatically calibrating the mass spectrum and identifying primer extension products for base calling is commercially available (Applied Biosystems).

The practical upper limit to multiplexing the SNP Genotyping Assay is 12-15 primers. If greater multiplexing is attempted, primer quality, particularly of the larger primers becomes important. Larger primers typically contain a greater proportion of contaminants such as failure sequences, and these smaller contaminants may produce signals that overlap signals from extension products of smaller fragments. Also, primers should be under 30 bases long to maximise signal intensity.

4.6 Primer Quality

A great advantage of the SNP Genotyping assay is that it does not require labelled primers, so the primers are relatively inexpensive. Generally, primers do not have to be of high purity. For assays with a low degree of multiplexing, desalted or cartridge-grade oligonucleotides are generally of adequate purity. Multiplexed assays with more than six primers may benefit from HPLC or PAGE purified oligonucleotides to remove lower mass contaminates, such as synthesis failure fragments. Primers which have been overheated or exposed to excessive acid during synthesis may contain amounts of abasic depurination products (producing peaks ~110-150 Daltons lower than the mass of intact primer). Abasic peaks can usually be ignored (see example in Figure 6), except that in highly multiplexed reactions it is

possible for depurination peaks to overlap other peaks. In this case, the primers should be resynthesised. Note that trihydroxyacetophenone (THAP) matrix (see below) always causes some depurination of oligonucleotides during the MALDI process, so for highly multiplexed assays 3-hydroxypicolinic acid (3-HPA) matrix is the better choice.

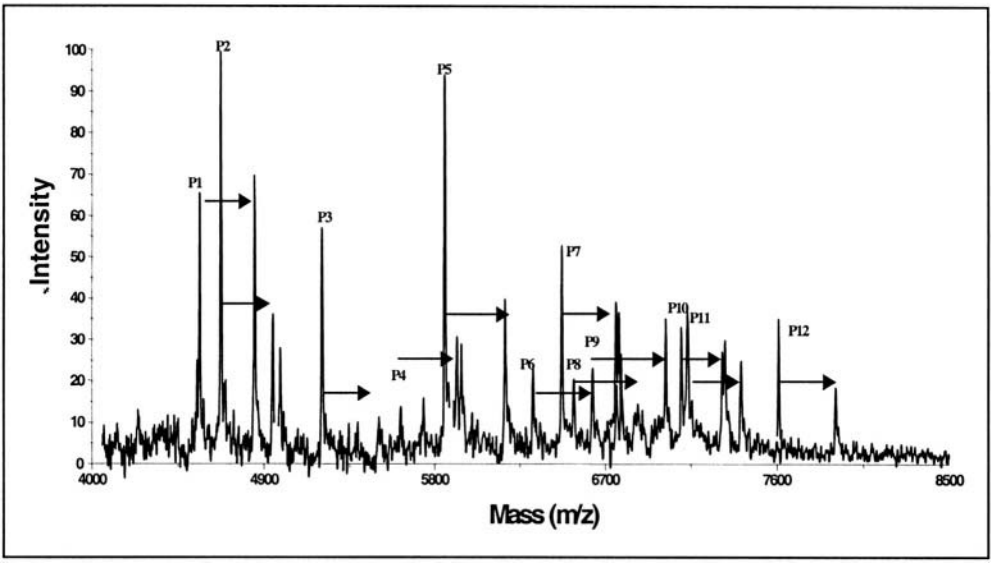


Figure 5. 12-Plex Mass-Tuned Genotyping Assay. A 12-plex multiplex PCR of human identity markers [20] was genotyped using a mixture of 12 primers [9] from 15 to 25 bases long.

5. THE SINGLE BASE EXTENSION REACTION

The SNP Genotyping Assay requires a thermostable, non-proofreading DNA polymerase that efficiently incorporates ddNTPs. Proofreading enzymes are to be avoided, because they do not produce any more accurate results, but instead cause extensive primer degradation [10]. *ThermoSequenase* DNA polymerase (Amersham/Pharmacia), and *AmpliTaq FS* DNA polymerase (Applied Biosystems) have both been validated for the SNP Genotyping Assay [10-14]. Recently, it was discovered that two other closely related polymerases work well: *Tth* DNA polymerase (Roche Molecular Systems) and *rTth* DNA polymerases (Applied Biosystems). These enzymes are less expensive than *ThermoSequenase* or *AmpliTaq FS DNA* polymerases, and both efficiently incorporate ddNTPs (in the presence of manganese ion). *Tth* DNA polymerase is employed in the Sequazyme™ Pinpoint SNP Typing Kit.

Reaction protocols suitable for *ThermoSequenase* and *AmpliTaq FS* DNA polymerases have been previously described in detail [9]; the protocols are similar for *rTth* or *Tth* DNA polymerase. Reaction mixtures typically included: 0.125 units/μl of *Tth* DNA polymerase, 0.5 μM of each genotyping primer, and 0.25 to 0.50 volumes of polished PCR product per volume of final reaction mixture. A typical reaction volume is 10 μl.

The recommended thermal cycling protocol (1 second 94° C – 2.5 minutes 37° C, for 25 thermal cycles, followed by a 4° C hold) does not require a third, higher-temperature primer extension step as commonly used in PCR. This low-temperature extension permits efficient extension of short and A+T rich primers. Thermal cycling 25 times provides a linear amplification of primer extension product, which may exceed the concentration of the PCR target sequence up to about 25-fold.

5.1. Desalting of Primer Extension Reactions

While the MALDI-TOF process is relatively forgiving for the presence of small amounts of salt, the bulk of contaminating sodium and potassium salts must be removed from the sample before analysis. This is most conveniently carried out by a simple reverse-phase desalting step [9], which in the past has been accomplished in pipette tips and can now be supported with 96-well filter plates. Samples can also be desalted by ethanol precipitation or by membrane float dialysis [15], although these techniques are less convenient for processing a large number of samples.

5.2. MALDI-TOF Conditions

Typically, 0.5 – 1 μl of desalted primer extensions are mixed with an equal volume of matrix solution and 0.5 μl of the mixtures dried onto a 384-well Teflon-coated MALDI sample plate. The most convenient matrix to use is trihydroxyacetophenone, or THAP, together with ammonium citrate [16, 17]. THAP crystallises evenly and yields a reproducible signal from position to position within a spot. However, THAP causes some depurination of the oligonucleotide primers during MALDI-TOF analysis, which is not generally a problem with multiplexed mixtures spaced at 2-base intervals. However, depurination can cause overlap problems in a multiplex reaction. The matrix 3-hydroxypicolinic acid, or 3-HPA, does not cause depurination [9, 17] and is recommended for complex mixtures. The MALDI-TOF analysis is generally carried out in positive ion mode, although negative ion mode produces similar spectra.

5.3. Determination of Bases Added to the Primer

The base or bases added to the primer are identified by comparing the mass differences between the unextended genotyping primer and the corresponding primer extensions with the calculated masses of the dideoxynucleotide base residues (Table 1). High mass accuracy is not required, and calibration of the mass

spectrometer is not important if both primer peaks and extended primer peaks are observed. This is because base identification is determined by mass difference between two closely spaced peaks rather than absolute mass. However, should a primer become fully extended, there will be no unextended primer to provide a mass for that calculation. This is not a problem if the mass spectrum is calibrated; in which case the mass difference can then be found by subtracting the *calculated* mass of the primer from the *measured* mass of the primer extension peak. The calibration can be based upon any other peaks of unextended primers in the mixture (it is extremely rare in a multiplex reaction for all the primers to become fully extended), or by addition of an internal standard oligonucleotide. If desired, the presence of unextended primer peaks in the mass spectrum can be assured by spiking some of the original primer mixture back into the sample after the primer extension reaction.

Table 1. Masses of ddNTP Single Base Extensions

Base	Daltons added
ddA	297.21
ddC	273.18
ddG	313.21
ddT	288.20

6. MODIFICATION OF THE SNP TYPING ASSAY TO SUPPORT ALLELE FREQUENCY DETERMINATION

Although MALDI-TOF is poorly quantitative in the absolute sense, it is good for relative quantitation. That is, when comparing peak areas of MALDI peaks of DNA of about the same size and base composition, the peak areas will accurately reflect the relative proportions of these substances. For example, in the case of a primer extended from a heterozygote target, the two different bases added to the primer will normally produce equal peak areas within a relative factor of +/- 10%.

In any individual, at any one SNP site, all individuals are either homozygous for wild type, homozygous for mutant, or heterozygous. The resolution of MALDI-TOF is adequate to resolve all six of the possible heterozygous biallelic primer extensions: A/C, A/G, A/T, C/G, C/T, and G/T. G and C, with a mass difference of 40 Daltons, are the most easily resolved. The most frequently occurring biallelic SNP genotypes are A/G and C/T, differing by 16 and 15 Daltons, respectively. The most difficult biallelic genotype to resolve is A/T, because A and T differ by only 9 Daltons.

Pharmacogenomic studies ultimately depend upon correlating allele frequencies to phenotypic traits, either within or between populations. Allele frequencies can be obtained by genotyping the individuals in that population, or by pooling the genomic DNA from a number of individuals and measuring the frequency of occurrence of an allele within the pool in a single measurement. Considerable time, labour and expense can be saved through determining an allele frequency through the single

measurement of allele frequency in a pool. Pooling can be accomplished by combining individual PCRs, or by forming a genomic DNA pool and performing a single PCR. By the second approach, assuming that both alleles amplify with equal efficiency, the measured allele frequency in the PCR product from the pooled genomic DNA should reflect the allele frequency of the original genomic DNA. The heterozygosity of a given variation can range up to 50%, and while there is no lower limit it is generally most productive to work with loci with a heterozygosity frequency above 1%.

As long as one is attempting to discriminate only between heterozygous and homozygous sample, there is adequate resolution by MALDI-TOF to measure peak areas of adjacent peaks of all possible biallelic loci, the most challenging being A/T, which differ in mass by only 9 Daltons. However, in a pooled sample in which the allelic ratio may be heavily biased towards one allele, it is not possible to accurately assign peak areas to the peaks corresponding to the two alleles. This problem can be addressed by substituting a modified dideoxynucleotide triphosphate in the primer extension reaction in place of one of the naturally occurring ddNTPs to increase the mass difference between the bases added and therefore their resolution. Ross *et al.* [19] recently described the use of FAM-labelled ddNTPs to accurately measure the minor allelic population down to about 2% of the major allele. Fei and Smith [15] screened twenty commercially available mass modified ddNTPs for this application. All of the mass modified ddNTPs enabled large mass shifts, typically between 600-800 Daltons, and *AmpliTaq FS* DNA polymerase effectively incorporated many. However, they reported evidence that all the tested mass modified bases reduced the detection efficiency of the extended primers, and that this needed to be considered in converting peak areas into allele frequencies.

Independently, we have found that biotinylated nucleotides are very efficiently incorporated and also do not appear to decrease MALDI-TOF signal. The terminators ddATP, ddGTP and ddUTP are incorporated into primer extension products about equally well as their non-biotinylated counterparts, although biotinylated ddCTP is incorporated significantly more slowly. This varied somewhat depending upon the DNA polymerase (results not shown here). In the case of *ThermoSequenase,* biotinylated-ddUTP was incorporated a little faster than unsubstituted ddUTP. This leads to the strategy that the six possible biallelic types can be clearly distinguished by typing each with a mixture of one unmodified ddNTP and one biotinylated ddNTP. As a practical matter, for all primer extensions involving C extensions together with another base, unsubstituted ddCTP should be used because C is the lightest base and biotinylated ddCTP is not efficiently incorporated.

As an example, a genotyping primer was designed for typing the ApoLipoprotein E (APOE) site 112, a T/C site encoding a cysteine-arginine interchange. Homozygote and heterozygote PCR products could be readily distinguished using a mixture of ddTTP and ddCTP, although primer extensions containing low amounts of the minor allele could not be fully resolved. However, a mixed sample containing only 10% of the rarer allele (i.e. 20% heterozygote), produced a spectrum with

completely resolved peaks corresponding to ddC (+273 Daltons) and biotinylated ddU (+666 Daltons) extensions (figure 6).

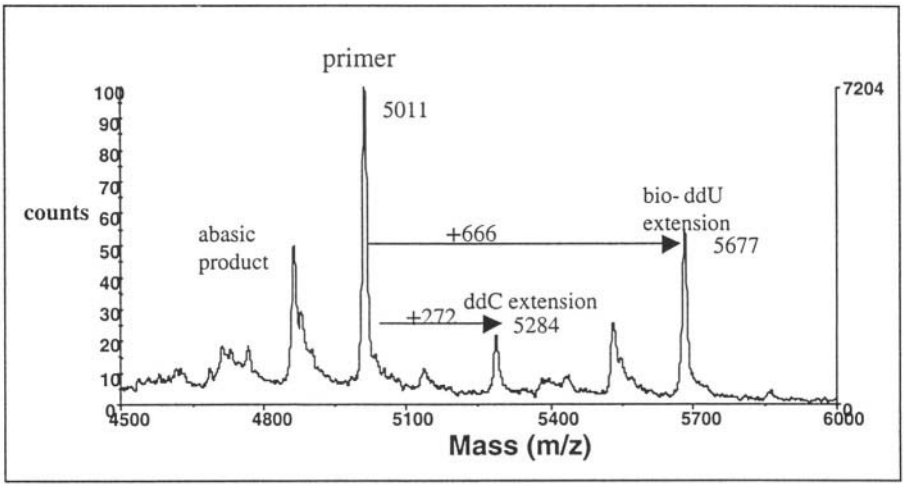


Figure 6. Allele Quantitation of APOE PCR Product with ddCTP and biotinylated ddUTP. Two PCR products were prepared, one from a heterozygote and one from a homozygote at site 112, a T/C polymorphic site in the ApoLipoprotein E gene. The two PCR products were mixed in a 1:5 heterozygote:homozygote ratio and typed with a genotyping primer of sequence 5'-TGGGCGCGGACATGGAGACC-3' with a mixture of ddCTP and biotinylated ddUTP.

A plot of the measured allele ratios (by peak areas), vs. the known allelic composition of pooled heterozygote/homozygote samples, showed a linear relationship between known allelic composition and measured peak area of the biotinylated ddUTP peak, with a slope of about 1.2. This reflects the slightly greater incorporation rate of biotinylated ddUTP to ddCTP (figure 7). Provided the PCR producing the sample to be typed is reasonably robust, it is possible to routinely measure minor allele peak areas down to about 1% of the area of the major allele peak. The slope determined using unmodified ddNTPs, in comparison, was about equal to unity, although it was difficult to accurately measure peak areas of the minor alleles employing unmodified ddNTPs when the minor allele was legs than about 20% of the total.

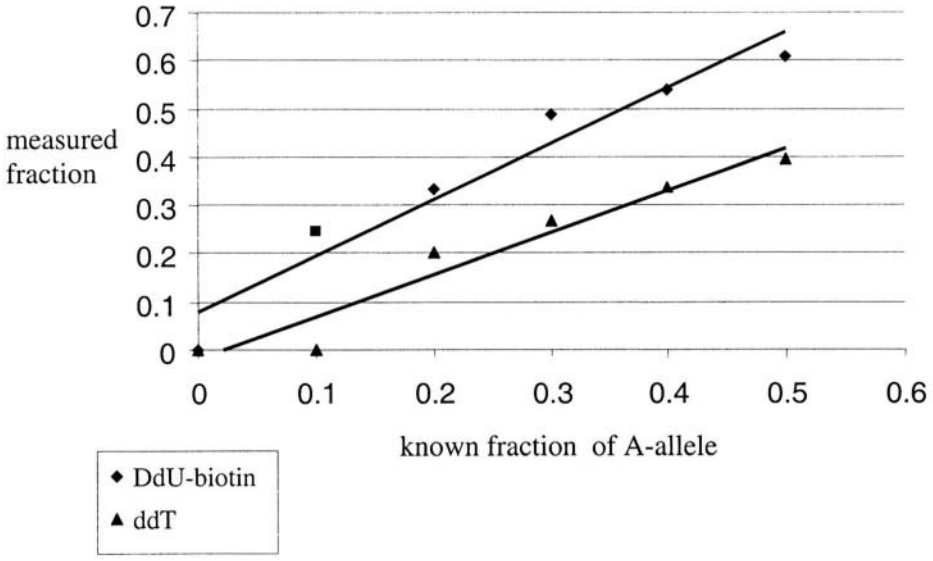


Figure 7. Allele Quantitation with Modified ddNTP. Mixtures of ratios of homozygous and heterozygous PCR products for the APOE E locus were prepared and genotyped with ddCTP and ddTTP (regular Sequazyme-PinPoint assay), or a mixture of ddCTP and biotinylated ddUTP. The measured fraction of A allele, as assayed by the peak area of fraction of ddT or biotinylated ddU incorporation, was plotted vs. the known fraction of the A allele in the mixture.

7. CONCLUSIONS

The Sequazyme Pinpoint assay has successfully typed hundreds of different loci in many different laboratories. Major advantages of the technique are that primer selection is extremely simple and a high percentage of selected primers successfully genotype. The primers need contain no special labels or groups, so they are inexpensive and easily obtainable. Because the PCR and primer extension protocols are solution-based and the extension protocol is universal, the assay does not involve any of the relatively high infrastructure costs of DNA chips. Because MALDI-TOF is largely artefact-free and high in resolution, the base-calling accuracy is extremely high. The ability to detect and quantitate low relative abundance of the minor allele in pooled samples is largely unique to mass spectrometry, due to its high resolution compared to electrophoretic, chromatographic, fluorescent, and isotopic techniques. The ability to process several thousand samples a day, together with the ability to internally multiplex, enables throughputs of about 20,000 alleles a day, suitable for most medium to high throughput genotyping laboratories. Future developments include further miniaturisation to lower reagent costs, faster lasers to support higher

throughput MALDI-TOF, and associated robotic systems to provide greater throughput.

8. REFERENCES

1. Cargill M, Alshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh j, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley GQ, Lander ES. *Nature Genetics* 22: 231, 1999
2. van Rensburg SJ, Potocnik FCV, deVilliers JNP, Kotze MJ, Taljaard JJP. *Annals of New York Academy of Science* 903: 200, 2000
3. Lutz CT, Foster PA, Noll WW, Voelkerding KV, Press RD, McGlennen RC, Kirschbaum NE. *Clinical Chemistry* 44: 1146, 1998
4. Hoogendoorn B, Owen MJ, Oefner PJ, Williams N, Austin J, O'Donovan MC. *Hum.Gen.* 104: 89, 1999
5. Nelson NC. *Critical Reviews in Clinical Laboratory Sciences* 35: 369, 1998
6. Greenwood AD, Burke DT. *Genome Research* 6: 336, 1996
7. Goelet P, Knapp MR, Anderson S. US Patent Number 5888819, 1999
8. Fahy E, Nazarbaghi R, Zomorrodi M, Herrnstadt C, Parker W, Davis RE, Ghosh S. *Nucleic Acids Research* 25: 3102, 1997
9. Ross PL, Hall L, Smirnov IP, Haff, LA. *Nature Biotech.* 16: 1347, 1998
10. Haff LA, Smirnov IP. *Biochemical Mass Spectrometry* 24: 901, 1996
11. Haff LA, Smirnov, IP. *Biochemical Mass Spectrometry* 1996, 901, 1998
12. Haff LA, Smirnov IP. *Nucleic Acids Research* 25: 3749, 1997
13. Haff LA, Smirnov IP. *Genome Research* 7: 378, 1997
14. Roskey M, Juhasz P, Smirnov IP, Takach EJ, Martin SA, Haff LA. *Proc. Natl. Acad. Sci. USA* 93: 4724, 1996
15. Fei Z, Smith LM. *Rapid Communications in Mass Spectrometry* 14: 950, 2000
16. Pieles U, Zurcher W, Schar M, Moser HE. *Nucleic Acids Research* 21: 3191, 1993
17. Zhu YF, Chung CN, Taranenko NI, Allman SL, Martin SA, Haff L, Chen CH. *Rapid Communications in Mass Spectrometry* 10: 383, 1996.
18. Li YCL, Cheng S, Chan T-WD. *Rapid Communications in Mass Spectrometry* 12: 993, 1998
19. Ross PL, Hall LR, Haff LA. *BioTechniques* 29: 620, 2000
20. Belgrader P, Marino, MM, Lubin M, Barany, F. *Genome Science and Technology* 1: 77. 2000

CHAPTER 3

MASSARRAY™: HIGHLY ACCURATE AND VERSATILE HIGH THROUGHPUT ANALYSIS OF GENETIC VARIATIONS

Hubert Köster
Sequenom Inc., 11555 Sorrento Valley R, San Diego, CA 92121, USA. Current address of the author: Villa Wellingtonia,CH-6917, Figino, Ticino, Switzerland. Email:hkoster1 @san.rr.com

1. INTRODUCTION

The small sequence variations between different human beings are not only responsible for differences such as eye or hair colour and attributes such as intelligence, musicality but also causing predisposition to genetic diseases and differences in drug response. The most frequent sequence variations are single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs). SNPs are position-specific DNA sequence changes, which occur with a frequency in the population of >1% (this number is somewhat arbitrary); a random mutation or a sequencing error is therefore not a single nucleotide polymorphism. Most of the SNPs are biallelic genetic markers, i.e. at the same sequence position two different bases could be found in the population with a frequency >1%. Since the DNA sequence is correlated to the protein sequence via the Genetic Code SNPs in coding regions (transcribed SNPs or cSNPs) are especially relevant to be associated with predisposition to disease, the efficacy of drug response and side effects due to differences in drug metabolism. Health care will see a revolution if we are able to understand how SNPs are associated with diseases such as arteriosclerosis, asthma, osteoporosis and cancer. It will also lead the way to individualised medicine, i.e. the development of drugs based on the molecular mechanisms (biochemical pathways) and the SNP-modified protein structure and function. It is our assumption that those drugs are not only more efficient in treating the appropriate genotype but also will significantly reduce drug side effects. In the future it can be envisoned that based on that knowledge first the specific SNP pattern (genotype) of a patient will be diagnosed followed by a treatment with the corresponding medicine. It will be the first time that diagnostics and therapy will work hand-in-hand on a rational basis for the benefit of the patient. This development will also pave the way to the necessary shift in medicine from therapy to prevention with the result of improved quality of life and reduction of costs within the healthcare system.

J.N. Housby (ed.), Mass Spectrometry and Genomic Analysis, 33-49.

It is believed that there are about 3 million SNPs in the human genome of which most of them are inconsequential; the estimate is that those SNPs which are of medical relevance are in the range of ten thousands and it may be that only a handful of coding SNPs (cSNP) are associated with a specific subtype of a disease.

If individualised medicine is to become reality at least two significant challenges have to be overcome: A technology has to be developed which allows to filter out of the millions of non-relevant SNPs those which are specifically associated with disease and a very accurate diagnostic/genotyping technology has to be created to diagnose patients for their personalised drug treatment.

2. MASSARRAY™ TECHNOLOGY

The MassARRAY™ technology has four fundamental components all of which are of equal importance. MALDI-TOF MS (Matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry) allows the analysis of DNA sequences directly (without using a label) through the determination of the sequence-specific molecular weight. Since molecular weight is an inherent physical property of a molecule, mass spectrometry has become, for small molecules, the gold standard in every analytical laboratory. In order for mass spectrometry to work, molecules have to be volatilised and ionised. Due to the low volatility and fragility of many biopolymers, standard mass spectrometry was not compatible until MALDI-TOF MS had been developed [1].

DNA fragments of SNP-differentiating and specific DNA sequences are generated through the MassEXTEND™ reaction in which a short oligonucleotide sequence (primer) is designed to selectively bind upstream to the SNP region to be analysed, serving as a primer for a DNA polymerase which enzymatically extends the primer in a highly reliable way downstream through the SNP region in the presence of deoxynucleoside triphosphates for elongation and dideoxynucleoside triphosphates for base-specific termination. Since in MassEXTEND™ at least one of the four nucleotides is only present as a terminator, the primer will be elongated until the first dideoxynucleotide is incorporated; it follows that the primer and all sequences resulting from the enzymatic primer extension must have different lengths. A separation step for educts and products is not needed since they are simultaneously separated by MALDI-TOF MS due to the differences in molecular weight. In contrasting technologies, which use labels, a separating step such as gel electrophoresis is essential and encumbered by all the various known artefacts, which are inherent to a separation step, based e.g. on the need that the analyte molecules must interact with the separating media for the separation to occur.

Since most SNPs are biallelic genetic markers, MassEXTEND™ can generate, for a given SNP, only three distinctly different primer extension products, i.e. three distinguishable different molecular weight values: Either both chromosomes in a diploid genome have, at the SNP position, the same sequence; in this case we can have the “normal” or the “disease” allele which must produce different extension products (homozygote normal/diseased ▶ only one albeit different extension

product, i.e. one peak in the mass spectrum) or both chromosomes differ at that SNP position (heterozygote ▶ resulting in two extension products – the homozygous normal and homozygous diseased one – i.e. two peaks in the mass spectrum). Figure 1 schematically describes the principle of SNP analysis by the MassEXTEND™ reaction. The sense template strand, derived from a PCR reaction, which in this instance is immobilised on a solid support via streptavidin-coated paramagnetic beads bound to its biotin moiety (B); as a representative example, a mixture of dTTP, dATP, dGTP and ddCTP are used. In template 1 (the G allele) the primer is extended by four bases; in case of a G▶A transversion (the A allele, template 2) the primer is extended by seven bases. In case of a heterozygote both peaks will be displayed. It is obvious

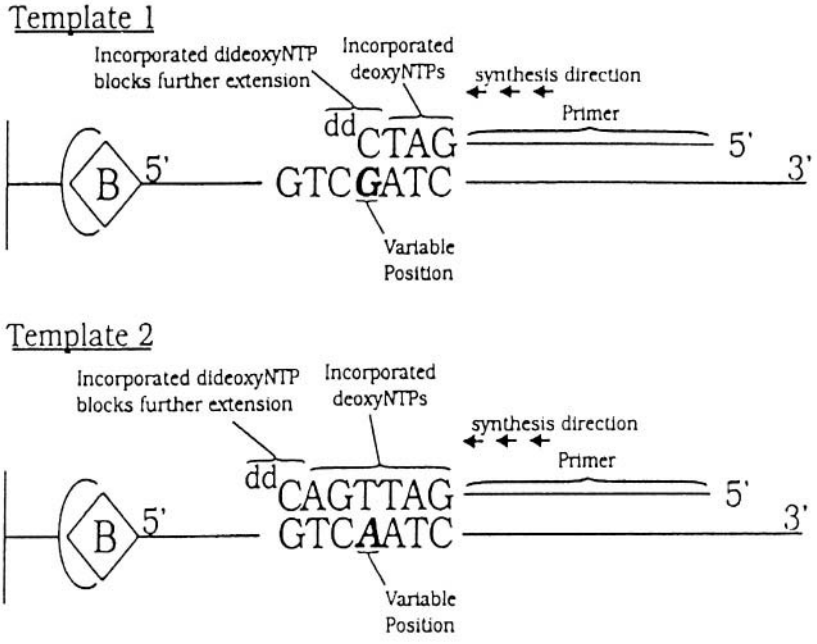


Figure 1. Schematic representation of the MassEXTEND™ reaction for SNP analysis.

that the MassEXTEND™ reactions generate analytical signals of very low complexity providing unambiguous analytical results and allowing for a high degree of multiplexing.

In addition to MALDI-TOF MS and the MassEXTEND™ reaction there are two additional fundamental features to MassARRAY™ which are important in order to enter the high throughput mode on an industrial scale, "Industrial Genomics™". For industrial genomics to become a reality, SNP analysis has to be transformed into an automated and cost-effective process. This is accomplished by using the SpectroCHIP™ – a sample holder for the mass spectrometer, in which minute amounts (femtomoles/nanoliters) of the analyte solution are co-crystallised with the matrix in a chip format i.e. positioned in a spatially resolved manner. The analyte is spotted onto the SpectroCHIP™ using either a robotic system employing piezoelectric pipettes or a pintool. In the latter case transfer is very fast and takes only minutes for up to 384 positions on a SpectroCHIP™. Ten of such SpectroCHIP™s are placed in a cartridge and run in one batch in a SpectroREADER™ mass spectrometer i.e. 3840 samples are analysed during one run. The nanoliter amounts of the analyte form miniaturised spots of a few hundred microns on the silicon wafer surface of the SpectroCHIP™, which, for the first time, allowed for fully automated signal acquisition by MALDI-TOF MS [2, 3]. Since the analyte is presented to the pulsed laser beam in a highly homogeneous form and is ablated over the whole spot surface, the analytical performance is significantly improved compared to standard sample preparation. Moreover, high mass accuracy and resolution is achieved with only a few laser shots thereby speeding up the analytical procedure. Importantly, it also enables determination of relative concentrations of molecules of similar size by comparing the respective mass spectral peak areas under normalised conditions.

The last important ingredient of MassARRAY™ technology is the signal acquisition and processing software, SpectroTYPER™, which not only refines the signals without changing their physical origin but also automatically translates the molecular weight data into relevant SNP/genotype information. This means that, in principle, no spectra need to be visualised since the genotype information is collected in Excel format, on a spreadsheet, and can be analysed on an Oracle based platform for its medical relevance.

The use of the SpectroCHIP™ leads the way to the industrial scale-up of the genotyping process through automated signal acquisition and interpretation and allows for significant cost-effectiveness – probably the most important feature of an industrial process - due to the miniaturised amounts of sample needed to run the MassARRAY™ process.

3. METHODOLOGY OF MASSARRAY™ TECHNOLOGY

The application of MassARRAY™ technology for the analysis of nucleic acid fragments has special demands for achieving high mass accuracy and resolution in order to meet the challenges for industrial genomics and diagnostics. One problem relates to the polyanionic properties of nucleic acids. In standard biochemical reactions such as PCR, LCR, primer extension reactions and DNA sequencing, many different salts are employed which lead to cation heterogeneity on the polyanionic

backbone. This jeopardises any analytical procedure with the extraordinary power of mass spectrometry since it causes peak broadening on the high molecular weight (right) side of the parent ion peak and therefore makes it impossible to determine the molecular weight with necessary accuracy. Additionally, DNA is most frequently obtained double stranded, which may lead to peak broadening since the molecular weights of the sense and antisense strands are different and in most cases not resolved. We addressed both problems by using solid supports, namely streptavidin-coated paramagnetic beads and biotinylated DNA fragments for the sample preparation, with subsequent ion exchange with ammonium salts, on the immobilised DNA fragments, prior to mass spectrometric analysis [4]. It is noteworthy that the non-covalent biotin-streptavidin interaction was stable under MALDI MS conditions whilst the hybridised antisense strand could be desorbed and analysed. It turned out that this was a very valuable concept, by combining standard nucleic acid biochemistry with MALDI-TOF MS it provides single stranded nucleic acid fragments with ammonium ions as the only, if not predominant, counter-ions. Ammonium is the optimal counter-ion since during laser-induced desorption it transforms the DNA in the gas phase into the protonated form [5]. A comprehensive study of the stability of the streptavidin-biotin system using paramagnetic beads with ammonium hydroxide, revealed that conditions could be found in which, predominantly, the hybridised antisense strand was released, followed by the release of the biotinylated sense strand from the beads for MALDI-TOF MS analysis [6]. This allowed the capture of a PCR product, where only one of the primers was biotinylated, on the magnetic streptavidin beads enabling each of the strands, or the double strand, available for MALDI-TOF MS analysis after ammonium cation exchange. This increased the flexibility of the system tremendously.

In DNA analysis one has to cope with another problem: the DNA molecule compared to RNA is more fragile i.e. under MALDI-TOF conditions the N-glycosidic bond is cleaved at A, G and, remarkably, also to some extend at C residues with subsequent breakage of the polyphophate diester backbone resulting in fragmentation with separate peaks or at least peak broadening on the low molecular weight (left) side of the parent ion peak. For A and G the fragmentation process starts with the protonation of N7 in the purine ring. Performing the PCR reaction by substituting the normal dATP and dGTP with the corresponding N7 deaza analogs we could obtain double stranded DNA demonstrating much better MALDI-TOF MS performance. In addition to achieving higher mass accuracy it was found that such modified DNA desorbed more efficiently, i.e. sharp and high intensity peaks were obtained with significantly less laser shots [7].

Further performance improvements in analysing nucleic acids could be achieved by the introduction of 3-hydroxypicolinic acid as matrix [8] and the introduction of delayed extraction in a linear time-of-flight mass spectrometer [9]. If, for MALDI Fourier transform mass spectrometry, the molecular weight range in analysing nucleic acid fragments could be extended further this type of MALDI MS would become of significant value due to the extraordinary resolution possible [10, 11]. In order to reach the sensitivity level necessary for MALDI-TOF MS analysis an amplification step has to be incorporated into the sample preparation process for

most applications. The method of choice is certainly PCR (see above). However other methods such as ligase chain reaction (LCR) can also be used [12]. For DNA sequencing reactions PCR and sequencing were combined by using two complementary DNA polymerases [13].

The next generation of solid supports in conjunction with MALDI-TOF analysis of nucleic acids is the transformation to a chip format, in which oligonucleotides or PCR products are covalently attached to a silicon wafer and all subsequent biochemical reactions and the ion exchange are performed on the chip surface [14, 15]. This is the next enhancement of the SpectroCHIP™, adding to the universal sample platform (as described above) customised versions for specialised applications.

4. DIAGNOSTIC APPLICATIONS OF MASSARRAY™ TECHNOLOGY FOR ANALYSIS OF DNA SEQUENCE VARIATIONS

A mutation in the human Factor V gene plays a major role in causing thrombosis (Factor V Leiden mutation: R506Q). The mutation (a G►A transversion) that changes the amino acid arginine (codon: CGA) to glutamine (codon: CAA) is located at codon 506. The site of this mutation is selected and amplified by PCR in furnishing a DNA of 224 base pairs of which a piece of the antisense sequence is shown below:

5' ...GTATTCCTCGCCTGTCCAGGGATCTGCTCTTACAGATTAGA...3'
3' ←CGGACAGGTCCCTAGACGAGAAT 5'

3' TAAGGAG----------(I) 5'
3' TAAGGAA----------(II) 5'

If the primer is extended to produce a sequence with a molecular weight of 9312 Da the, "normal" allele with G, at the underlined position, is present (product I); if, however, the molecular weight value found is 9296 the mutant with the disease relevant predisposition is diagnosed (product II). Figure 2 displays the result from one patient. The raw mass spectrum shows only two main peaks, one at 7082 Da representing some of the unextended primer (theoretical mass: 7082 Da), the second peak with a mass of 9297 demonstrates the presence of the Factor V Leiden mutation in that patient. The small peak on the high molecular weight side at 9485 Da reflects the presence of a citrate molecule (adduct formation from ammonium citrate in the matrix crystals), the other small peak to the left side of the main peak is a depurination peak (see discussion above). In this experiment the MassEXTEND™ reaction was performed with a mixture of dATP, dGTP and ddTTP. The presence of dCTP would have been without influence, however if one wants to keep the option open to simultaneously detect any new mutations in that region the presence of all dNTPs with the exception of the fourth one as ddNTP would be an advantage. In the case of a heterozygote one would have seen two peaks of 16 Da mass difference

(9912 vs. 9296 Da) which represents the mass difference between an A and a G residue. Figure 3 gives an example of a G/A heterozygote from a different site of the human genome. A 16-mer primer is being extended with a mixture of dATP, dGTP and ddTTP. The theoretical mass of primer-dAdAddT is 5828 Da and that of primer-dGdAddT is 5844 Da respectively. In Figure 3 raw data are presented; the peaks are clearly separated and the experimental mass values are in excellent agreement with the theoretical ones. This experiment demonstrates that MassARRAY™ has been developed to a state where these small mass difference are easily differentiated and detected. If one only plans to diagnose in the experiment described in Figure 2.

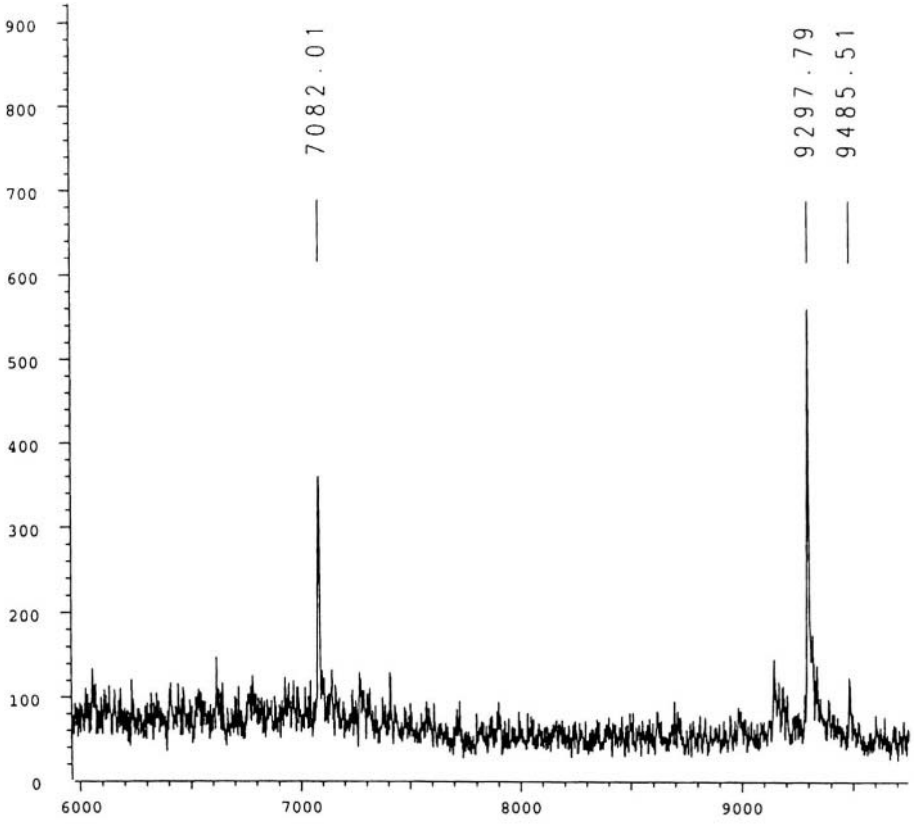


Figure 2. MassARRAY™ raw spectrum of the MassEXTEND™ reaction extending a 23-mer primer with a mixture of dATP, dGTP and ddTTP in a 7 base extension reaction (theoretical mass: 9296 Da, found 9297 Da) diagnosing the disease allele, i.e. A (Gln) instead of G (Arg) at the second position of codon 506 of the factor V gene.

which of the two alleles are present (i.e. leaving out the option to simultaneously analyse the downstream region) than a mixture of only dATP and ddGTP would

result in a much broader mass difference since in the case of the normal allele the primer will be elongated by one nucleotide (ddGTP) and in case of the disease mutation by three nucleotides (dAdAddG) respectively. The mass difference is so large (626 Da) that no error can be made in distinguishing these two alleles In the following, some more applications of MassARRAY™ for the analysis of sequence variations in nucleic acids including SNPs and STRs are briefly summarised since all of the experimental details are in the public domain.

For the detection of pathogens such as bacteria and viruses and also significant chromosomal aberrations (e.g. deletions, insertions, rearrangements) the molecular weight of a PCR product, or, even better, a nested PCR product, is sufficient to unambiguously diagnose the presence or absence of infectious DNA or chromosomal aberrations. As an example, we were able to detect 10 hepatitis B (HBV) viral genomes in 100 µl of blood [16]. In a subsequent study we confirmed, with full agreement (100%), the results of standard procedures (PCR and gel electrophoresis and hybridisation with a ^{32}P-labeled hybridisation probe) for all uninfected, control, and for the infected patient samples. In a third group of patients which were anti-HBc positive, standard methods could not detect any viral DNA whereas MassARRAY™ technology had proven that about 50 % of those patient samples still contained viral HBV DNA [17].
In general terms, MassARRAY™ turned out to be a fast, highly sensitive and non-radioactive method for the detection of PCR products without the use of error-prone and laborious gel electrophoresis or hybridisation with labelled probes.

Numerous assays have been developed to detect DNA sequence variations such as SNPs by the MassEXTEND™ reaction (formerly named PROBE = PRimer OligoBase Extension) [18]. A few selected examples should be discussed here.

Mutations in the cystic fibrosis transmembrane regulator gene (CFTR) are medically relevant in causing cystic fibrosis (CF). The diagnostic products are generated by MassEXTEND™ reactions with a mixture of three dNTPs and the fourth only present as a terminating ddNTP. The single diagnostic primer which bound upstream of the variable site was extended through the region of interest by a DNA polymerase until the first ddNTP was incorporated. The molecular weight of the extension product correlated with the composition of the variable site. Five CF mutations, from two sites of exon 11, were analysed in a biplex reaction (two diagnostic primers used simultaneously). In addition, assays were developed for the identification of three common alleles of the polyT tract at the intron 8 splice site of the CFTR gene [19]. The results were unambiguous and highly reliable. Due to the diagnostic importance of CF many different assays were developed such as reverse dot blots, amplification refractory mutation systems (ARMS), sequencing-by-hybridisation (SBH), oligo ligation assay (OLA), genetic bit analysis (GBA) and solid phase mini-sequencing. These allele-specific assays need high hybridisation/annealing stringency for high quality results which make the development of robust multiplex assays almost impossible. Especially, the detection of false positive and false negative results would have a detrimental effect on the diagnostic value of such methods. MassEXTEND™ is a non-allele-specific assay

and due to the low complexity of the signals, significant multiplexing is possible without jeopardising data quality. Multiplexing is highly desirable for CF diagnostics since there are more than 600 mutations known in the CFTR gene which have different frequencies in different ethnic groups. In order to achieve a significant diagnostic level as much as 50 to 100 mutations need to be analysed per patient. This is a significant challenge for any technology, however, addressable by MassARRAY™ technology.

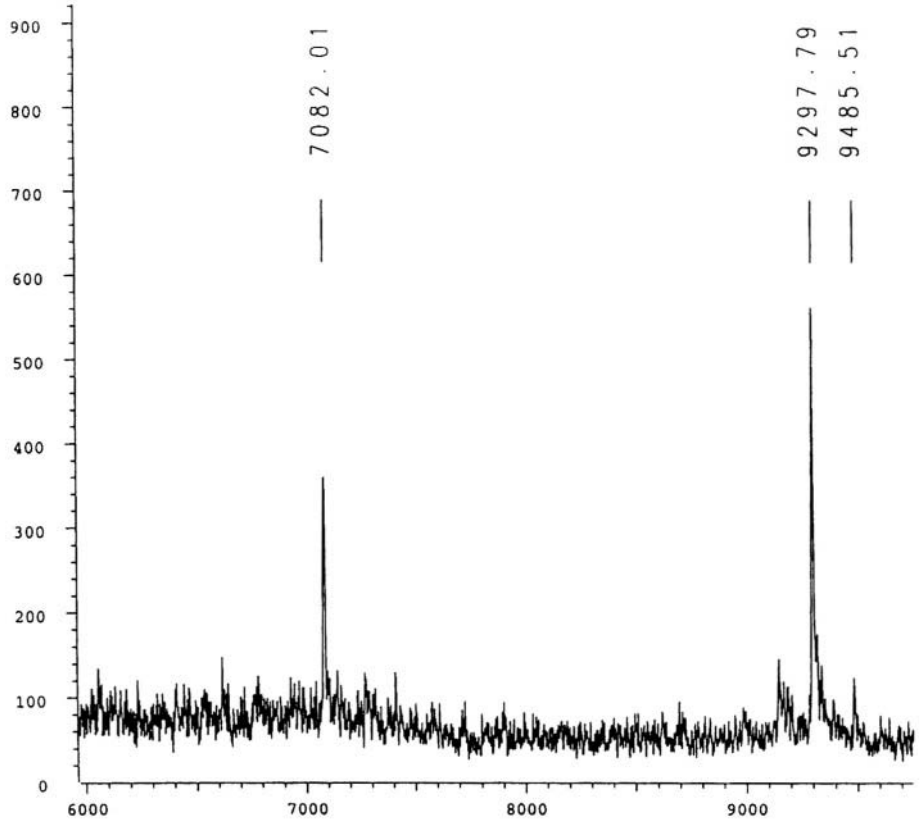


Figure 3. MassARRAY™ raw spectrum of a heterozygote (G/A) as diagnosed by a MassEXTEND™ reaction extending a 16-mer primer with a mixture of dATP, dGTP and ddTTP to primer- dAdAddT (5828 Da) and primer-dGdAddT (5844 Da).

The RET proto-oncogene located on chromosome 10 codes for a transmembrane tyrosine kinase. Mutations at codon 634 are directly associated with multiple endocrine neoplasia type 2A and medullary thyroid carcinoma. If children are diagnosed to carry such a mutation, cancer can be prevented by removal of the thyroid gland. Accuracy and reliability of the diagnostic assay is critical here as in

many instances; consider the consequences of false positive and false negative results [11].

Apolipoprotein E is a plasma protein involved in lipid metabolism; its genetic polymorphisms consist of three common codominant (ε2, ε3, ε4) and several less common alleles. The polymorphisms are within codon 112 and 158 of the 299 amino acid long protein. The ε2, ε3 and ε4 alleles contain Cys/Cys, Cys/Arg and Arg/Arg respectively and are associated with different and distinct medical conditions such as familial hyperlipoproteinaemia type III, ischaemic heart disease and Alzheimer's disease. The importance of a reliable and accurate diagnostic assay is obvious. This is the reason why almost any available technology has been applied to develop diagnostic assays which allow differentiation between all of the different alleles. MassEXTEND™ allows one, with only two specific diagnostic primers binding upstream of condon 112 and 158 respectively and an appropriately designed mixture of dNTPs and ddNTP, to simultaneously determine the important ε2/ε3, ε3/ε3, ε3/ε4 and ε4/ε4 genotypes. The mixture of triphosphates is designed such that each variable region results in unique extension products which can easily be detected by MALDI-TOF MS [20]. An important improvement of the primer extension reaction is also introduced with this study. The MassARRAY™ performance was significantly improved by performing the MassEXTEND™ reaction in a temperature-cycled mode; this lead to a dramatic increase of the extension products and allowed a reduction in the amount of sample necessary for the diagnostic assays.

While SNPs as biallelic genetic markers are especially important today for unravelling the associations of genetic variations within coding regions with disease, drug efficacy and drug tolerance, for genetic linkage and positional cloning studies multiallelic markers such as short tandem repeats (STRs) or microsatellites provide much higher efficiency because they are highly informative. The use of di-, tri- and tetranucleotide repeats has dramatically improved the establishment of genetic linkage maps. We successfully employed the MassEXTEND™ reaction to many STRs. For example, the AluVpA DNA marker (ATTT repeats) within intron 5 of the interferon-α receptor gene will be discussed [21]. Due to the relatively large molecular weight differences of the extension products with different repeat units, detection of most microsatellites by MassARRAY™ was easy. By varying the dNTP/ddNTP mixtures second site mutations within the repeat alleles could be detected which remained undetected by gel electrophoretic sizing methods. Thus, using MassEXTEND™ not only the exact number of repeats can be detected but at the same time second site mutations are revealed thereby significantly increasing the polymorphism information content. This is important for applications such as statistics-based gene mapping, cancer diagnostics (loss of allelic heterozygosity) and forensics.

The most sensitive analytical method is, of course, DNA sequencing; it is at the same time the biggest challenge for a mass spectrometric-based technology since the resolution and mass accuracy decreases with DNA fragment length so that currently Sanger sequencing ladders, generally, cannot be read beyond 100 nucleotide units. Although until now not the method of choice for *de novo* DNA sequencing it can

become an attractive alternative for diagnostic sequencing in which, generally, only DNA fragments of far less than 100 nucleotides need to be analysed. Indeed MassARRAY™ has shown to be suitable for diagnostic sequencing [22]. The advantage of using MassARRAY™ for sequencing became obvious. Frequent sequencing artefacts in Sanger sequencing using fluorescent labelling and gel electrophoresis, such as pausing or stuttering of the polymerising enzyme, could be easily detected. In the case of a "hole" in the sequencing ladder the molecular weight values of the flanking ladders allowed to directly calculate the missing and correct base. The accuracy, reliability and robustness of MassARRAY™ are important features for a diagnostic sequencing approach. By primer walking strategies we were able to sequence four exons (5 to 8) of the p53 gene [23]. The p53 gene codes for a tumour suppressor protein. Mutations in the region of exon 5 to 8 are associated with the formation of many different types of cancer. Since spontaneous mutations appear at random sites MassEXTEND™ cannot be applied. In those cases, also applicable for the BRCA1 and BRCA2 genes as well as for the AIDS virus and the MHC and HLA regions, only diagnostic sequencing can be used to reveal the sequence variations. With the encouraging results presented it can be envisioned that with some improvements in technology these important diagnostic challenges can be met by MassARRAY™ technology.

5. APPLICATION OF MASSARRAY™ FOR CONFIRMATION AND VALIDATION OF SINGLE NUCLEOTIDE POLYMORPHISMS

In the previous section the utilisation of MassARRAY™ for diagnostic applications has been discussed; this section will briefly focus on the technological challenges involved in our path to determine the association of SNP patterns with subtypes of diseases, i.e. to filter those relevant SNPs out of the millions of SNPs which are completely unrelated to causing disease or influencing the efficacy or tolerance of drug response.

The importance of this approach is demonstrated by the formation of The SNP Consortium (TSC) in 1999 in which 13 multinational companies and leading academic institutions are collaborating to discover 300,000 SNP markers covering the human genome. Until now about 100,000 SNPs have been disclosed in the public domain (http://snp.cshl.org/). Since these SNPs have been found *in silico* i.e. by mining the public DNA sequence databases, it is questionable whether all of them are real SNPs or e.g. sequencing errors. The real effort therefore starts after the *in silico* discovery of a SNP in that it has to be confirmed and validated which means that at first a sequencing error (or other experimental/computing artefacts) can be excluded (confirmation) and secondly that it is polymorphic i.e. occurs with a significant frequency in the population at that sequence position and is not a random mutation (validation). To understand the scope of such a project the tasks ahead of the TSC will be discussed. To confirm and validate the different 300,000 SNPs means that 300,000 individual assays have to be designed, developed and optimised. Most of the other SNP analysing technologies need days, if not weeks, for the development of a new optimised, i.e. robustly working, assay. If the development of

an assay could be done in one day and assuming a 200-working-day-year this would mean that the development of 300,000 assays will last for 200 years! An extension of the bioinformatics tools developed for MassARRAY™, however, has lead to an assay design software which fully automatically designs 10,000 and more assays per day with a success rate to work as a robust assay the first time it is tried in the laboratory of more than 80%. Why is this possible with MassARRAY™ ? Due to the fact that molecular weight values are the signals of an MassEXTEND™ reaction the whole assay design process could be done on the computer since one is dealing only with numbers of known components (the triphosphates, the primers, the extension products etc.). In addition, the MassEXTEND™ reaction can be designed in a way that the results are analytically error-free because the mass differences between educts and possible products can be modulated so that they exceed the mass accuracy and resolution achievable today with MassARRAY™ by at least one order of magnitude. Finally, MassEXTEND™ eliminates many artefacts seen in technologies which use labels and separating steps such as gel electrophoresis and high performance liquid chromatography (HPLC) because it is a direct method based on detecting molecules through their individual molecular weights; these are the main reasons for the high success rate of computer-aided assay design.

While assay development for such a project already has industrial dimensions, the confirmation and validation process of SNPs is even more demanding. In order to get to a sufficient level of statistical significance that a sequence variation at a single sequence position is really polymorphic (frequency of about 10% in a population) it is believed to be reasonable that the same site in the genome is analysed in at least 100 different individuals of the same ethnic background and geographic ancestry. This means that for 300,000 SNPs, 30 million SNPs (or genotypes) have to be analysed. With a technology which can perform 100,000 genotypes per day such a study (assuming again a 200-working-day-year) would still last for one and a half years. In these industrial scale applications, high throughput has a different meaning than in low scale R&D applications: The analytical accuracy of the technology is becoming of extraordinary importance. For instance, a technology which is 99.9% accurate will still produce 30,000 erroneous results! This means that when one is moving from the R&D environment into industrial scale applications (Industrial Genomics™) only the most accurate technology can be used: In an industrial world errors are no more affordable! This becomes clear if we are going one step further in our path to individualised medicine. The most challenging part is to find associations between candidate SNPs to subtypes of diseases. In this instance, one has to analyse samples from patients suffering from the disease and compare them with the same number of samples coming from individuals who are not affected by the disease in order to filter out those SNPs which are associated with the predisposition of the disease of interest. Common understanding is that at least 1000 samples in each group have to be analysed to reach a sufficient level of statistical significance that a specific SNP is indeed correlated with the disease in question. Taking again the 300,000 SNPs (now we assume that these are all confirmed and validated) have to be analysed individually in 2000 samples (1000

from the disease, 1000 from the healthy population). This translates into running 600 million genotypes! With a technology which analyses 100,000 genotypes per day such a study would take 30 years (assuming again a 200-working-day-year)! Needless to say the costs of such a study taking today's levels make it completely cost-prohibitive. Since the goal is to discover a selected handful of coding SNPs associated with different subtypes of a given disease, the use of a technology which is 99.9% accurate will jeopardise such a study since one has to deal with potentially 600,000 errors! If the initial 300,000 "confirmed and validated" SNPs contain potentially 30,000 erroneous results (see above) the situation is even worse.

The solution to this dilemma is MassARRAY™ technology. Its robustness and accuracy and opportunity for a high degree of automation allows, instead of analysing SNPs in individuals, to pool, lets say, the DNA of 100 individuals and to perform the site-specific PCR and MassEXTEND™ reaction in a pooled sample. In the example given above, one now only needs to analyse 2 pools (1 disease and 1 healthy pool) of 1000 individuals each i.e. for the 300,000 SNPs 600,000 genotypes or, better, allelotypes need to be analysed. With a technology able to analyse 100,000 genotypes/allelotypes per day the whole study would take now one week instead of 30 years. Needless to say that the costs are dramatically reduced compared with conventional technologies.

There is still some way to go but MassARRAY™ has the potential to be the technology of choice for industrial scale SNP confirmation, validation and association studies to enable pharmacogenomics, thereby leading the way to disease specific SNP assays for the diagnosis of disease predispositions and drug efficacy and tolerance and providing the molecular basis for the development of individualised medicine. A first significant step has been made towards that goal: In a collaboration with the Laboratory of Population Genetics at the US National Cancer Institute out of a total of 10,243 *in silico* SNPs 9,115 SNP assays have been automatically designed. A total of 6,404 SNPs were experimentally proven to be polymorphic of which 3,148 were previously unknown. The DNA of 94 individuals had been pooled in those experiments. SNP allele frequencies were determined with an average accuracy of ±1.6% based on the rare allele being present in the sample pool with a frequency of 10%. This collaboration generated the largest validated coding SNP collection to date [24]. It is noteworthy that the experimental part from start to finish had been performed in less than four weeks.

6. CONCLUSIONS

The completion of the Human Genome Project generated a raw DNA sequence of the human genome. We are now at the beginning of an era in which we have to extract the huge information embedded in the sequence and to transform it into knowledge about the molecular mechanisms of disease predisposition and development, drug efficacy and tolerance. The amount of work which lies ahead is colossal. In addition genomic experiments are highly interdisciplinary and complex. With respect to genotyping, traditional methods are technically demanding, inaccurate, costly and in most cases difficult to automate. In short they are

unsuitable for the industrial scale of the tasks ahead. We have to understand that in order to capitalise on the DNA sequence information towards developing a health care system which shifts therapy to prevention with the benefit of improved quality of life we are leaving the territory of traditional R&D and entering the industrial world. We termed this new era Industrial Genomics™. In general terms an industrial process has to be very flexible, accurate, reliable and robust, allowing high throughput through simplicity and automation and all together leading to a cost-effective process. Traditional technologies and procedures are unsuitable for industrial scale applications. MassARRAY™ is the technology that has been designed to meet the challenges of Industrial Genomics™.

The following proprietary features summarise the advantages of MassARRAY™ over conventional technologies:

Accuracy: MALDI-TOF MS allows the analysis of molecules directly – no labels to be introduced and quality-controlled, no separation step with image processing necessary. Instead, direct measurement and simultaneous separation through the molecular weights of the molecules. Fast signal acquisition in one-thousandth of a second and repetitive and fully automated sample processing in 3-5 seconds; the signal is an electronic signal and directly processed in the computer for data analysis. MassEXTEND™ is not hybridisation-dependent (no mismatch artefacts!); internal standards lead to high reproducibility.

Flexibility: MassEXTEND™ is a universal process for the investigation of genetic variations and allows the analysis of a region (not just a single nucleotide position) with one single primer; to detect known or even unknown point mutations, deletions, insertions, reversions, repeats and substitution polymorphisms in that region. Different alleles are precisely identified including triallelic systems. For industrial applications most important: Fully automated assay design is possible to generate ten thousands of assays per day with a laboratory success rate of more than 80%.

Throughput: Due to high accuracy and resolution, multiplexing is possible thereby analysing more than one sample simultaneously. Most parts of the analytical process are currently automated allowing one batch to analyse 3840 samples (uniplex or multiplex) in 3 to 4 hours. With a pentaplex reaction and two 8 hour-shifts close to 100,000 genotypes/allelotypes could be analysed per day with the state of the technology development today.

Automation: MassARRAY™ allows automated assay design, automation of sample preparation and data analysis. Despite the high degree of automation the process remains flexible, simple and highly accurate.

Cost-effective: Nanomolar quantities (femto moles of analyte) in nanoliter volumes are sufficient through the miniaturised SpectroCHIP™ format and the sensitivity and accuracy of MassARRAY™. Multiplexing and the high success rate of computer-designed assays add to the cost-effectiveness of the MassARRAY™ process. The high degree of automation and simplicity of the operation reduces labour costs. In the industrial world it has to be understood that high throughput should not be achieved by sacrificing accuracy. Subsequent repetitive experiments

or interpretation of false data based on inaccurate results are simply too costly to be affordable, i.e. at an industrial scale accuracy and costs are directly correlated. Most important is the possibility that MassARRAY™ allows in the analysis of pooled samples with the same high accuracy and reliability. This dramatically reduces costs and basically is the necessary prerequisite to enable industrial scale projects in the genomics area.

7. MATERIALS AND METHODS

General: MALDI-TOF mass spectrometry has been performed firstly using a Finnigan Vision 2000 mass spectrometer with a reflectron, later also with a PerSeptive Voyager and recently with the Sequenom-Bruker array mass spectrometer (SpectroREADER™). The latter are both linear (with ~ 1 m field-free drift region) TOF instruments with delayed extraction. In all cases a nitrogen laser with nanosecond pulses (wavelength 337 nm) has been employed. DNA segments of the genome were amplified via PCR (Taq DNA polymerase, "HotStar",Qiagen) using one biotinylated primer. Purification was performed on streptavidin-coated paramagnetic beads and the non-biotinylated strand denatured. For the MassEXTEND™ reaction the primer was hybridised to the immobilised strand, the appropriate mixture of dNTPs and ddNTPs added and extended using Thermosequenase (Amersham Pharmacia Biotech). For generation of increased amounts of the extension product a temperature cycling program was applied the exact condition depending on the respective individual experiment. When the SpectroREADER™ was used nanoliter quantities of the analyte were spotted onto the surface of a SpectroCHIP™ with either a piezoelectric or pintool robot. The analyte was spotted onto 200μ spots on the SpectroCHIP™ prefilled with matrix crystals (mixture of 3-hydroxypicolinic acid and ammonium citrate). After recrystallisation of analyte and matrix the SpectroCHIP™ with either 96 or 384 positions was ready for automatic readout by the SpectroREADER™. The dedicated SpectroTYPER™ software for e.g. baseline correction, peak identification and automatic genotype calling processed the signals.

PCR reaction: Typically 5.0 μl of the dNTP/ddNTP mixture (each 2 mM), 1.0 μl (10 pmol) 5'-biotinylated forward primer and 1.0 μl (25 pmol) unbiotinylated reverse primer, 5.0 μl genomic DNA (5 ng/μl), 0.2 μl HotStar Taq DNA polymerase (5 units/μl) and 5.0 μl reaction buffer for HotStar Taq DNA polymerase (10X) in a total volume of 50 μl were used. Temperature program: 30 – 50 cycles of typically 20 sec denaturation at 95°C, 30 sec annealing at 56°C and elongation at 72°C for 30 sec.

Preparation of template: Typically for a reaction 15-μl streptavidin-coated paramagnetic beads (Dynal) are washed twice with 200 μl each of the B/W buffer (Dynal), suspended in 15 μl B/W buffer and added to the PCR reaction mixture. After incubation for 30 minutes at room temperature, the supernatant was removed through magnetic separation and the beads incubated with 50 μl 100 mM aqueous NaOH for 5 minutes at room temperature to denature the non-biotinylated DNA

strand. The supernatant was removed and the beads washed twice with 50 µl each of 10 mM Tris/HCl, pH 8.0.

MassEXTEND™ reaction: Typically 2.0 µl (20 pmol) primer, 1.5 µl (50 µM end concentration) of the appropriate dNTP/ddNTP mixture and 0.5 µl (2.5 units) of Thermosequenase, 1.5 µl reaction buffer for Thermosequenase in a reaction volume of 15 µl. This reaction mixture is added to the streptavidin-coated paramagnetic beads harbouring the immobilised single stranded template DNA and the extension reaction performed using a temperature program: Typical 1 min at 80°C, 3 – 10 cycles of 10 sec annealing at 40°C, 5 sec elongation at 72°C and at the last cycle 3 min at 72°C. The supernatant is then magnetically removed and the beads are washed 2-3 times with 50 µl 10mM Tris/HCl, pH 8.0 and the extension product denatured by employing 5.0 µl of 50 mM aqueous ammonium hydroxide solution and incubation for 4 min at 60°C. Supernatant is removed by magnetic separation, transferred to an Eppendorf tube and shaken for 20 min (open tube) at room temperature. Samples should be stored frozen if not directly used. Aliquots are transferred to the SpectroCHIP™ and analysed as described before.

Specific experimental conditions and procedures are given in the publications listed under references.

8. ACKNOWLEDGEMENTS

The author thanks Dipl.-Chem. Silke Atrott for making available figures 2 and 3

9. REFERENCES

1. Karas M, Hillenkamp F. *Analytical Chemistry* 60: 2299, 1988
2. Little DP, Cornish TJ, O'Donnell MJ, Braun A, Cotter RJ, Köster H. *Analytical Chemistry* 69: 4540, 1997
3. Little DP, Braun A, O'Donnell MJ, Köster H. *Nature Medicine* 3: 1413, 1997
4. Tang K, Fu DJ, Kötter S, Cotter RJ, Cantor CR, Köster H. *Nucleic Acids Research* 23: 3126, 1995
5. Nordhoff E, Ingendoh A, Cramer R, Overberg A, Stahl B, Hillenkamp F, Grain PR *Rapid Communications in Mass Spectrometry* 6: 771, 1992
6. Jurinke C, van den Boom D, Colazzo V, Lüchow A, Jacob A, Köster H. *Analytical Chemistry* 69: 904, 1997
7. Siegert CW, Jacob A, Köster H. *Analytical Biochemistry* 243: 55, 1966
8. Wu KJ, Steding A, Becker CH. *Rapid Communications in Mass Spectrometry* 7: 142, 1993
9. Wiley WC, McLaren IH. *Review Scientific Instruments* 26: 1150, 1953
10. Li Y, Tang K, Little DP, Köster H, Hunter RL, McIver RT Jr. *Analytical Chemistry* 68: 2090-2096, 1996
11. Little DP, Braun A, Darnhofer-Demar B, Frilling A, Li Y, McIver RT Jr., Köster H. *Journal of Molecular Medicine* 75: 745, 1997
12. Jurinke C, van den Boom D, Jacob A, Tang K, Wörl R, Köster H. *Analytical Biochemistry* 237: 174, 1996
13. Van den Boom D, Ruppert A, Jurinke C, Köster H. *Journal of Biochemical and Biophysical Methods* 35: 69, 1997
14. O'Donnell MJ, Tang K, Köster H, Smith CL, Cantor CR. *Analytical Chemistry* 69: 2438, 1997
15. Tang K, Fu DJ, Julien D, Braun A, Cantor CR, Köster H. *Proceedings of the National Academy of Sciences USA* 96: 10016, 1999

16. Jurinke C, Zöllner B, Feucht HH, Jacob A, Kirchhübel J, Lüchow A, van den Boom D, Laufs R, Köster H. *Genetic Analysis Biomolecular Engineering* 13: 67, 1996
17. Jurinke C, Zöllner B, Feucht HH, van den Boom D, Jacob A, Polywka S, Laufs R, Köster H. *Genetic Analysis Biomedical Engineering* 14: 97, 1998
18. Köster H, van den Boom D, Braun A, Jacob A, Jurinke C, Little DP, Tang K. *Nucleosides & Nucleotides* 16: 563, 1997
19. Braun A, Little DP, Köster H. *Clinical Chemistry* 43:7: 1151, 1997
20. Little DP, Braun A, Darnhofer-Demar B, Köster H. *European Journal of Clinical Chemistry & Clinical Biochemistry* 35: 545, 1997
21. Braun A, Little DP, Reuter D, Müller-Mysok B, Köster H. *Genomics* 46: 18, 1997
22. Köster H, Tang K, Fu DJ, Braun A, van den Boom D, Smith CL, RJCotter, Canto Cr. *Nature Biotechnology* 14: 1123, 1996
23. Fu DJ, Tang K, Braun A, Reute D, Darnhofer-Demar B, Little DP, O'Donnell M, Cantor CR, Köster H. *Nature Biotechnology* 16: 381, 1998
24. Buetow KH, Edmonson M, Macdonald R, Clifford R, Yip P, Kelley J, Little DP, Strausberg R, Köster H, Cantor CR, Braun A. *Proc. Natl. Acad. Sci.* USA 98: 581, 2001

CHAPTER 4

THE GOOD ASSAY

S. Sauer, D. Lechner, IG. Gut
Centre National de Genotypage, Batiment G2, 2 rue Gaston Cremieux, CP 5721, 91057 Evry Cedex, France. Tel: 0033-160-878-359; Fax: 0033-160-878-383; Email: ivogut@cng.fr

1. INTRODUCTION

After the completion of the Human Genome Sequencing Project, genome research is bound to focus on genome variation studies. In one project a consortium of pharmaceutical companies and academic research groups is creating a dense genome-wide map of single nucleotide polymorphisms (SNPs) that will serve as markers for genotyping experiments. In another project 10.000 genes are being scanned for SNPs that will be of use for the identification of disease susceptibility variants. By adopting systematic and large-scale approaches to such studies, geneticists are hoping to gain valuable insight into the relationships between genes, gene variants and phenotypes that would never have been predicted based on previous biological knowledge.

Single nucleotide polymorphisms have an average frequency of approximately 1 per 1.000 base pairs in the human genome [1]. They are considered to be the ideal marker for the dissection of complex traits using association studies and linkage disequilibrium mapping. Efficient and economic genotyping technologies with high-throughput are very sought-after for these projects, where it is likely that cohorts with several tens of thousands of individuals will be analysed for their genetic variability [2]. Pharmacogenomics using SNP analysis may become a powerful tool for medicine, where, depending on a particular genotype, a patient will get the most appropriate medication [3]. First examples of this sort of application have been demonstrated [4, 5]. Many methods for SNP genotyping, like microarrays, gel-based and plate-reader based assays have been described [6]. None of these match the qualities of rapid analysis and accuracy of mass spectrometers, which are well suited for genotyping large cohorts.

2. SNP GENOTYPING BY MALDI

Already in the late 1980s, with the introductions of matrix-assisted laser desorption/ionisation time-of-flight mass spectrometry (MALDI) and electrospray ionisation mass spectrometry (ESI), it was recognised that the mass spectrometer might be an ideal tool for the analysis of DNA, proteins, and peptides. Initially, it

J.N. Housby (ed.), Mass Spectrometry and Genomic Analysis, 50-65.

was thought that MALDI mass spectrometry would replace fluorescence DNA sequencers in genome sequencing projects. After an initial burst of publications in the early 1990s dealing with the analysis of oligonucleotides by MALDI, a lot of research groups abandoned the field when it was realised that progress was not quite as easy as anticipated. For this reason the adaptation of DNA preparation to MALDI has taken significantly longer then expected and the human genome sequencing project was completed without using mass spectrometry.

Nevertheless, over the last ten years, mass spectrometric methods used for the analysis of DNA have come of age, so that now several procedures for the elucidation of DNA sequence variation are ready for integration into the post-genome sequencing era. Several of the protagonists of these developments have contributed chapters to this book. It was almost universally acknowledged that one of the major problems of DNA analysis by mass spectrometry was to achieve sufficient sample purity. Rather than to look for solutions for sample purification we decided to look for the reason for the sensitivity of DNA to impurities.

3. HOW TO IMPROVE THE ANALYSIS OF DNA BY MALDI

MALDI was initially applied to the analysis of proteins, especially in the emerging field of proteomics (described in another chapter of this book). DNA is significantly more difficult to analyse because of its chemical structure and properties [7]. The main problem in analysing native DNA by MALDI is its negatively charged sugar-phosphate backbone. Mass spectrometry relies on a molecular charge for analysis. With native DNA, the phosphate residue provides a site of negative charge in solution and each DNA molecule carries as many negative charges as phosphate residues. The affinity of the phosphate residues for alkali counterions, such as sodium and potassium, and even metal counterions, is high, but not high enough to result in a complete saturation. These ions interfere with the ionisation process, by inducing adducts and thereby limiting the signal intensity and quality [8]. The use of ammonium counterions in MALDI is a well-established method to counteract ion affinities [9]. In solution ammonium exits as a NH_4^+ counterion, whereas in the gas-phase NH_3 is readily lost, leading to a reduced counterion structure. However, ammonium ions introduce a degree of suppression to the desorption process. This results in a dramatic decrease of analytical resolution and sensitivity. Nowadays stringent purification procedures are applied to counteract these problems. These include magnetic bead separation and reversed-phase column purification that tend to be cumbersome in high-throughput applications. Another predicament is the acid instability of DNA. Sample preparation is done with acidic matrices and acidic conditions are encountered in the desorption/ionisation process. In the gas-phase, DNA can readily fragment with harsh matrices. A detectable degree of depurination has been observed for larger DNA products [6, 10]. Replacing purines by 7-deaza-analogues is one approach to prevent DNA from depurinating [11, 12]. A second approach to improve DNA in MALDI is the use of ribonucleotides containing 2'-OH groups that stabilise the gas-phase ion [13]. In a third approach it was found that the replacement of phosphate protons from native DNA backbones by alkyl groups

significantly improved the behaviour of the molecule in the MALDI process [14, 15].

The optimisation of the MALDI process consists of identifying the right matrix and preparation method for an analyte. How matrices function in MALDI is not well understood. The chemical structure of DNA is complex and its interaction with a matrix during the desorption/ionisation process eludes investigation. Only empirical findings progressed the method. It was observed that DNA analysis by MALDI was very inefficient [15], for example 100 times more DNA has to be used in a preparation to achieve a similar signal intensity comparable to peptides.

Our idea for rendering DNA amenable to analysis by MALDI focuses on the difference in analysing oligonucleotides and peptides. While most peptides are formally uncharged, DNA carries as many negative charges as phosphate bridges. Charges were neutralised by replacing phosphate groups by phosphorothioate groups and alkylating them. The efficiency of alkylation of regular phosphate groups is low, but a selective and quantitative alkylation is achieved with phosphorothioate groups. Furthermore it was known that the addition of a positive charge-tag to peptides changed their desorption behaviour [16]. Therefore the addition of a positive charge-tag with subsequent removal of all charges from the phosphorothioate backbone bridges was implemented. The concept of this was to generate a product with a defined charge state, thus relying on the matrix for desorption, but not for ionisation. Using this approach, there was a 100-fold increase of detection efficiency, equaling the detection efficiency of peptides [17]. The same result was observed when all but one backbone-bridge were neutralised and the DNA product thus carried a single negative charge (-1 charged DNA product) [18].

α-Cyano-4-hydroxy-cinnamic acid methyl ester turned out to be the ideal matrix system for DNA compounds with either one positive or one negative charge. It is the methyl ester of α-cyano-4-hydroxy-cinnamic acid, the most commonly used matrix for peptide analysis. In contrast to other matrices it has a significantly higher pK_a of around 8. This means that this matrix does not support the protonation of the DNA products during sample preparation. Its absorption maximum perfectly matches the emission wavelength of an N_2 laser, which is the most commonly used laser in MALDI mass spectrometers. In contrast, matrices used for protein and peptide analysis typically have a very low pK_a. Standard DNA matrices, like 3-hydroxypicolinic acid (HPA, the most common matrix for DNA analysts) have slightly acidic pK_a's around 4. One of the most striking observations with α-cyano-4-hydroxy-cinnamic acid methyl ester is that native DNA cannot be analysed with this matrix [17]. Use can be made of this discriminative behaviour as the selectivity of this matrix is towards singly charged DNA compounds. There is little difference in ionisation efficiency in negative or positive ion mode analysis of singly charged oligonucleotides with this matrix [18].

There are two common matrix preparation methods, thin-layer and dried droplet preparation. For thin-layer preparations the matrix is applied to the MALDI target plate in a volatile solvent, such as acetone. The solvent spreads and evaporates

immediately, leaving a thin layer of small matrix crystals. The analyte is dispensed onto the thin-layer in a solvent that does not dissolve the matrix. Analyte molecules are built into the surface of the matrix. For dried droplet preparations a matrix solution is mixed with an analyte solution and then spotted onto the MALDI target plate. Dried droplet preparations tend to give "sweet spots"- certain positions on the preparation give better results than others. This makes these preparations difficult to use in automated processes. Due to the uneven height of dried droplet preparations the mass calibration can also be unstable. MALDI analysis is based on the determination of the time-of-flight of an ion. Variable height of the matrix preparation results in a shift of the starting position which affects the time of flight. This can easily conclude in a few Daltons mass variation.

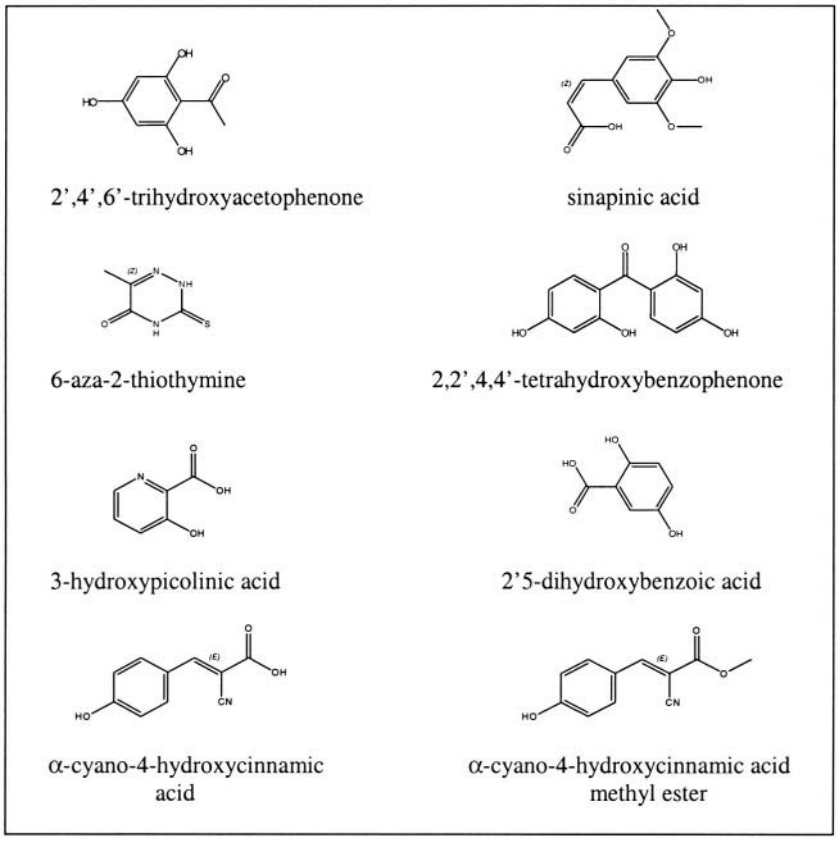


Figure 1. Chemical structures of common matrices for the MALDI analysis of DNA, proteins and charge-tagged DNA molecules.

In contrast, thin layer preparations give less spot-to-spot variation, better mass accuracy and resolution. Thin-layer was used with α-cyano-4-hydroxy-cinnamic

acid for peptide analysis, while DNA analysis preferably was done with HPA in a dried droplet preparation. Using α-cyano-4-hydroxy-cinnamic acid methyl ester with thin layer preparation for modified DNA gives a significant improvement of the reproducibility of the sample preparation in an automated set-up, while simple liquid handling systems can be used to deposit samples onto the MALDI target plate.

A prerequisite for the application of any DNA analysis method is that a specific DNA product can be made. For example, for a genotyping method allele-specific products have to be generated. Preferably this is done enzymatically, as enzymes can provide high specificity. α-S-dNTPs are substrates that can replace dNTPs in template-directed primer extension reactions with DNA polymerases [19]. Thus phosphorothioate DNA can be generated enzymatically. Complementary primer sequences containing phosphorothioate bridges in place of phosphate bridges anneal to template DNA with ample specificity and many DNA polymerases will readily extend them.

4. PRINCIPLES OF THE GOOD ASSAY

For the GOOD assay [20] the know-how of sensitivity enhancing chemistry for DNA analysis by MALDI was integrated into a molecular biological preparation procedure for allele-specific products. The two chemical modifications that needed to be integrated were the introduction of a charge-tag and charge neutralisation of the product DNA backbone. The GOOD assay starts with a PCR. PCR actually fulfils two functions, first it generates a sufficient amount of template for the allele-specific processing, and second it reduces the sequence complexity of the template decreasing the risk of mispriming. There are very few SNP genotyping procedures that do not require the amplification of a stretch of DNA from a genomic template by PCR prior to the actual allele distinction. The only published PCR free method is the Invader assay, which in turn has the drawback that it requires large amounts of genomic DNA [21].

As the third step of the GOOD assay is a primer extension with a specifically tailored set of α-S-dNTPs and α-S-ddNTPs, dNTPs of the PCR have to be removed. This is done enzymatically by degradation of the dNTPs in the second step by shrimp alkaline phosphatase (SAP).

A primer extension reaction is used to generate allele-specific products. A primer is chosen upstream of the SNP that is to be genotyped. Primers can be placed on either strand of the PCR product. They are synthesised with functionalities that will result in the final products being +1 or –1 in net charge. Primers with chemical modifications are shown in Figure 3.

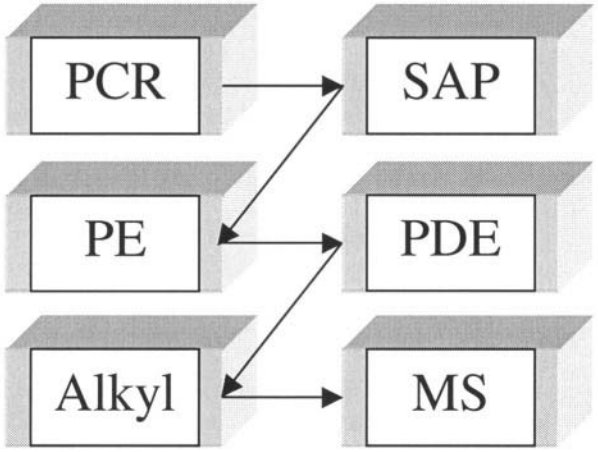


Figure 2. Flowchart of the GOOD assay. SAP stands for shrimp alkaline phosphatase digestion, PE for primer extension, PDE for phosphodiesterase, Alkyl for alkylation and MS for MALDI mass spectrometry.

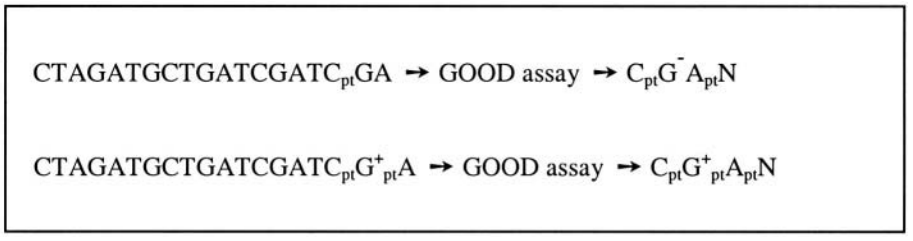


Figure 3. The primer is elongated during the GOOD assay by an allele-specific primer extension. For the negative ion-mode version of the GOOD assay the primer carries a phosphorothioate modification at the second bridge from the 3'end. For the positive ion-mode version a positive charge-tag carrying base is at the second base from the 3'end. This base is bracketed by two phosphorothioate bridges. The unmodified part of the primer is removed by 5'phosphodiesterase digestion, while the phosphorothioate bridges inhibit further degradation.

The last three bases at the 3'end of the primer are connected with two phosphorothioate bridges for positive ion-mode analysis (Figure 4). The middle base has an amino-modification that allows the attachment of the positive charge-tag. For the negative ion-mode, the second and the third base from the 3'end of the primer are connected by a phosphorothioate bridge. The last bridge is a regular phosphate group that carries the negative charge at the end of the GOOD assay. The

phosphorothioate groups fulfil two functions. They are quantitatively charge neutraliseable by alkylation and they inhibit the complete digestion of the primer in the fourth step of the procedure. For the primer extension reaction (step 3) primers are added together with a specifically selected set of α-S-dNTPs and α-S-ddNTPs. These substrates are readily accepted by a number of DNA polymerases. Their addition to the primer results in the formation of further phosphorothioate bridges. Typically, we position the primers immediately next to the position of the SNP and extend only with α-S-ddNTPs. This reduces the complexity of results of multiplexes and gives the most homogenous signal patterns.

The fourth step of the GOOD assay is the digestion of a large part of the primer with a 5'-phosphodiesterase (Figure 4). The primer does not contribute to the information content of the allele-specific product. Therefore it can be removed and the size of the products to be analysed can be significantly reduced. Thus the molecular weights of the products are shifted into a range where the detection sensitivity and resolution of the mass spectrometer is best. An alternative would be to use primers that have only phosphorothioate bridges. We found that these sorts of primers have significantly worse annealing properties. They are more expensive and generally the quality of these primers is poor.

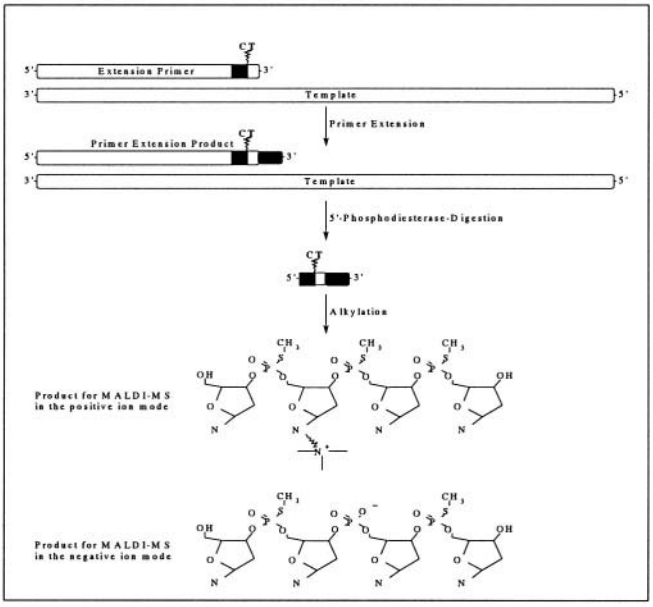


Figure 4. Principle of the GOOD assay. "CT" symbolises the charge-tag, "N" a DNA base and N^+ a quarternary ammonium.

In the fifth step of the GOOD assay the phosphorothioate bridges are charge neutralised by an alkylation reaction (Figure 4). The reaction conditions are chosen such that the selectivity of the reaction is optimal for the addition of methyl groups to phosphorothioate bridges, while no alkylation of the bases takes place. The addition of the alkylating agent also results in the reaction mixture separating into two phases. The allele-specific products are in the upper aqueous phase. The upper phase is sampled and diluted in a sample preparation solution. From there samples are transferred onto a MALDI target in a thin layer preparation and analysed. The product masses lie in a mass range of 1000 to 3000 Da. A crucial advantage over all other DNA analysis methods using MALDI detection is that no purification is involved in the GOOD assay.

The GOOD assay is a single tube procedure. The five reaction steps are done without transferring the samples to new reaction vials. Reagents are simply added to the reaction and the samples either placed in an incubator or thermocycler. This makes it amenable for automation.

The main benefit of the GOOD assay is that the introduction of sensitivity enhancing chemistry circumvents the need to purify and concentrate the products prior to MALDI analysis, making it very economical for SNP genotyping.

5. VARIATIONS OF THE GOOD ASSAY

The GOOD assay with negative ion-mode detection is easily accessible. The required primers can be obtained from most manufacturers of oligonucleotides. The GOOD assay with positive ion-mode detection, on the other hand, gives options for further derivatisations of the product oligomer.

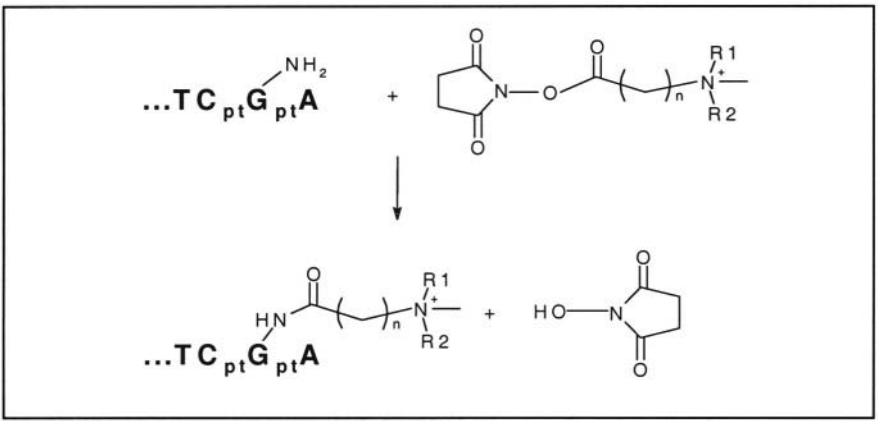


Figure 5. The charge-tag reaction of an amino-modified primer (3'-sequence in bold) and a charge-tag reagent (usually 6-trimethylammoniumhexyryl-N-hydroxy-succinimidyl-ester). For increasing the potential of multiplexing different charge-tag reagents can be used, with R1 and R2 being ethyl instead of methyl groups and n=4,5,6,8. Currently three charge-tag reagents with masses 158 Da, 144 Da, and 130 Da are available from Bruker Saxonia GmbH, Leipzig, Germany.

A major quality of mass spectrometers is that they can be used to analyse many different compounds in one experiment. In practice this means the mass spectrometer, as a multi-channel detector, can identify the alleles of many different SNPs at once. We have chosen to use the GOOD assay in multiplexes (simultaneous preparation and analysis of different SNPs) with interleaved allele products. As the products generated by the GOOD assay are usually four bases long this only gives 28 different base compositions (CCCC, CCCT, CCCA,...GGGG). A requirement for establishing a multiplex reaction is that any two alleles of the SNPs of the multiplex do not have the same base composition. This severely limits the possibilities for multiplexing. But with the possibility of further modifications a larger panoply of masses can be generated in the positive ion-mode.

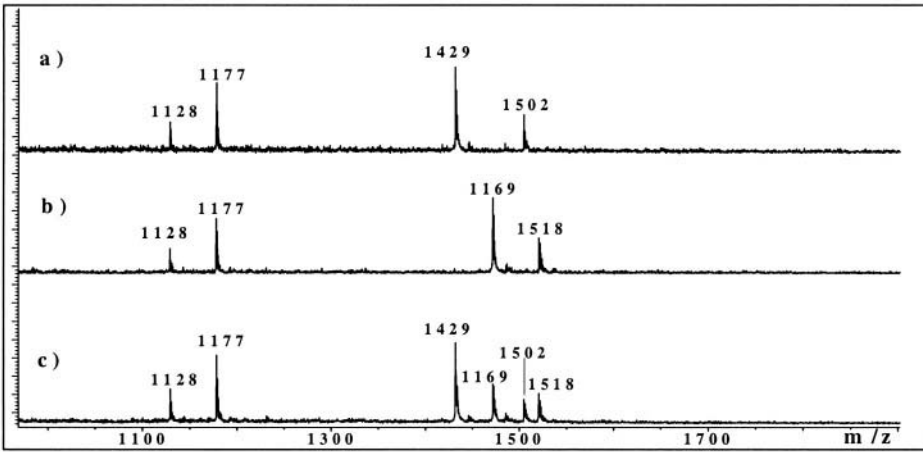


Figure 6. Two SNPs in the interleukin 6 gene at position 565 (A/G) and 987 (C/G) are analysed simultaneously. Masses with 1128 and 1177 m/z are residual modified primers. In a) DNA homozygous for A (product sequence AG^{CT}GA, 1502 m/z) at position 565 and homozygous for C at position 987 (product sequence TG^{CT}CC, 1429 m/z) is shown. In b) DNA homozygous for G (product sequence AG^{CT}GG, 1518 m/z) at position 565 and homozygous for G at position 987 (product sequence TG^{CT}CG, 1469 m/z) is shown. In c) heterozygous DNA for both SNPs was analysed.

For the GOOD assay in the positive ion mode version, amino-modified phosphoramidites A^{NH2}, G^{NH2}, C^{NH2}, and U^{NH2} are integrated into the extension primers [22]. Their introduction into the primers already increases the options for multiplexing. This can be further increased by the addition of charge-tags with different masses. There is a choice of different charge-tags. The amino-modified phosphoramidates, charge-tag reagents and the mentioned matrix for the GOOD assay are available from Bruker Saxonia GmbH (Leipzig, Germany).

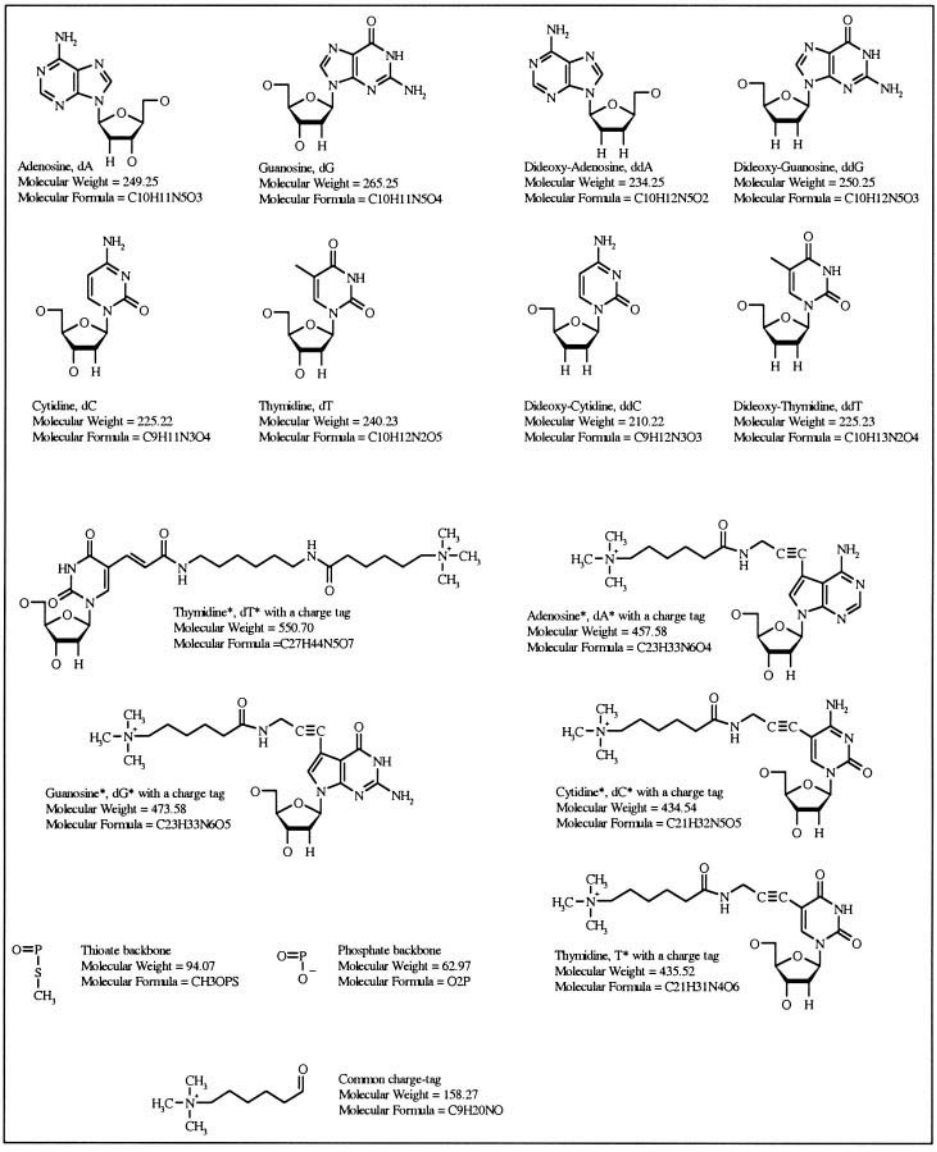


Figure 7. All compounds of the GOOD assay and their respective masses are shown. The amino-modified nucleosides are indicated with an asterisk and are charge-tagged. The most common charge-tag variant, deriving from 6-trimethylammoniumhexyryl-N-hydroxy-succinimidylester is shown at the bottom of the figure. Furthermore, the two possible linkages of the products of the GOOD assay, a phosphate group or an alkylated phosphorothioate group are displayed.

The ability to multiplex reactions is of great importance for reducing the cost per SNP genotype and analysis time. Multiplexing at the PCR level and of the primer extension reaction is currently under investigation. The quality of the enzymatic reactions in a multiplex assay depends on the SNPs that have to be combined. The surrounding DNA sequences have an profound influence on the quality of the PCR reaction.

In order to efficiently create multiplex experiments we have developed a software tool called the SNIP Machine [23]. With this software, the design of multiplex experiments is facilitated. The SNIP Machine determines the best combination of primers for the PCR and the following primer extension reactions, considering annealing temperature, sense of the DNA oligomer, mass separation (in the mass range 1000 to 3000 Da) and preparative cost. Because of the finite mass resolution of mass spectrometers a minimum mass separation of alleles is maintained. The SNIP Machine generates instructions for the required charge-tag reagents and extension nucleotides. Its output can be forwarded to an oligonucleotide provider. In Figure 6 an example of a duplex GOOD assay is shown. The signal to noise ratio allows automatic allele calling. In Figure 7 all compounds of the GOOD assay are shown.

6. MATERIALS AND METHOD OF THE GOOD ASSAY

Taq-DNA Polymerase, dNTPs and phosphodiesterase II (from calf spleen) were purchased from Roche Diagnostics (Mannheim, Germany). α-S-dNTPs and α-S-ddNTPs were provided by Amersham (UK). Thermosequenase and shrimp alkaline phosphatase were purchased from Amersham Buchler (Braunschweig, Germany). Standard chemical reagents were purchased from Aldrich (89555 Steinheim, Germany). α-cyano-4-hydroxy-cinnamic acid methyl ester, propargylaminomodified 7-deaza-2'-deoxyguanosine and 6-trimethylammoniumhexyryl-N-hydroxy-succinimidylester were provided by Bruker Saxonia GmbH (Leipzig, Germany).

Oligonucleotides for the PCR and primer extension reactions were synthesised and HPLC purified by MWG Biotech (Ebersberg, Germany). For the amplification of a fragment of the interleukin 6 gene primers 5'-CCTGGAGGGGAGATAGAGCTTCT-3' (forward) and 5'-GAGACGCCTTGAAGTAACTGCAC-3' (reverse) were used. 5'-CTGCACGAAATTTGA$_{PT}$G$^{CT}{}_{PT}$G-3' for the SNP 565 and 5'-CCCTAGTTGTGTCTT$_{PT}$G$^{CT}{}_{PT}$C-3' for the SNP 987 in the interleukin 6 gene were used as primers for the primer extension reaction.

Charge tagging: The amino functionality of the synthesised primers was used for attaching a positive charge tag (6-trimethylammoniumhexyryl-N-hydroxy-succinimidylester) according to Gut et al. [17] and Bartlett-Jones et al. [16]. The primers were dissolved in 1 % TE-buffer to 500 pmol/µl. 30 µl of this solution were mixed with 1.5 µl 2 M triethylammoniumhydrogencarbonate (pH 8,0) and 24 µl fresh 1% 6-trimethylammoniumhexyryl-N-hydroxy-succinimidylester solution. This

reaction mixture was incubated at 0 °C for 30 min. After the reaction mixture was lyophilised, the pellet redissolved and ethanol precipitated. The product was finally dissolved in 30 µl double distilled water. The efficiency of charge-tagging was monitored by MALDI. More than 95 % of each of the amino-modified oligonucleotides were converted into charge-tagged primers.

PCR: 1 µl of genomic DNA (~ 2 ng) were used as template for the PCR. 5 pmol each of the forward and reverse primer were mixed with a buffer (pH 8,8) containing 40 mM Trisbase, 32 mM $(NH_4)_2SO_4$, 50 mM KCl, 4 mM $MgCl_2$, 200 µM dNTPs and 0,5 U Taq Polymerase in a 10 µl final volume. The reaction was denatured 2 min at 95°C, then thermocycled 15 sec at 95°C, 30 sec at 67°C and 30 sec at 72°C 30 times. Finally the reaction was incubated at 72 °C for 4 min.

Shrimp alkaline phosphate (SAP) digestion: 0.5 µl (1 U/µl) of shrimp alkaline phosphatase were added to the PCR reaction and incubated for 1 h at 37°C. The enzyme was denatured for 10 min at 90°C.

Primer extension: 25 pmol of the charge-tagged primers were added together with 4 mM $MgCl_2$, 0,2 mM $MnCl_2$, 100 µM α-S-ddNTPs and 0.5 U Thermosequenase. The reaction volume was increased to 20 µl by the addition of water. An initial denaturing step 2 min at 95°C was done, followed by 35 cycles of 20 sec at 95°C, 1 min at 60°C and 30 sec at 72°C.

Primer Removal: 1 µl of a 0.5 M acetic acid solution was added to the processed primer extension reaction resulting in a reaction pH < 7. Then 2 µl of phosphodiesterase II that was previously dialysed against ammonium citrate (0.1 M, pH 6.0) were added and the reaction was incubated for 1 h at 37°C.

Alkylation reaction: 45 µl of acetonitrile, 15 µl of triethylammonium bicarbonate solution (pH 8,5) and 14 µl of methyliodide were added. The reaction was incubated at 40°C for 25 min. Afterwards 20 µl double distilled water were added. Upon cooling a biphasic system was obtained. The upper layer contained the products while the lower layer contained some of the reagents, for example detergents that were added to stabilise the enzymes used in this procedure. 20 µl of the upper layer were sampled and diluted in 45 µl of 40% acetonitrile. This solution was directly used to transfer the samples onto the matrix.

Sample preparation for MALDI analysis: The α-cyano-4-hydroxy-cinnamic acid methyl ester matrix was prepared by spotting 0.5 µl of a 1,5% solution in acetone onto the target and spotting 0.4 µl of a solution of the sample in 40% acetonitrile on top of the dried matrix thin-layer. 40% acetonitrile dissolves the surface layer of the matrix allowing for a concentrated incorporation of the analytes into the matrix surface.

Mass spectrometric analysis: Spectra were recorded on a Bruker Reflex III time-of-flight mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany). This mass spectrometer was equipped with a Scout MTP™ ion source with delayed extraction. Spectra were recorded in positive ion linear time-of-flight mode. Typical acceleration potentials were 18 kV. For delayed extraction, the acceleration potential was switched with a delay of 200 ns.

7. APPLICATIONS OF THE GOOD ASSAY

The GOOD assay can be applied for genotyping SNPs in a variety of scenarios. The majority of SNPs we developed were in the human genome. Systematic studies of SNPs in candidate genes in large cohorts (with DNA of several thousand individuals) have been completed. Furthermore, SNP genotyping for agricultural applications is of great interest. Markers for genetic fingerprinting of cattle have been developed using the GOOD assay. This could be applied for the traceability of animals. Another interesting application is the determination of a SNP associated with the susceptibility of pigs to porcine stress syndrome. Programmes applying SNPs for marker supported breading in plants are also becoming very popular. This is supported by the GOOD assay.

8. THE ISSUE OF DNA QUALITY

The quality of SNP genotyping experiments strongly depends on the quality of the DNA. In most cases DNA is extracted from whole blood using commercial kits. Because as little as 0.5-2.0 nanograms of genomic DNA suffice for genotyping a SNP by the GOOD assay, these DNA extractions tend to last quite long and the cost of extraction is spread over the number of SNPs that are genotyped. However, there are some applications where it is inconceivable to use expensive extraction procedures. Mainly these are for agricultural applications. In an example a system of tissue sample taking during ear-tagging of cattle has been adapted to a SNP production line. Tissue samples are digested with Proteinase K. This preparation is used straight for the GOOD assay without isolation of the DNA. The Proteinase K digest byproducts do not inhibit the GOOD assay [22].

9. PHYSICAL HAPLOTYPING BY THE GOOD ASSAY

Microsatellite genotyping benefits from the large number of alleles each marker can provide. Yet, they require analysis by gel-based methods and do not provide information about coding changes. SNPs on the other hand give only binary information. The information content of SNPs can be increased if the phase of multiple SNPs (haplotype) within a region is measured (Table 1). We have extended the GOOD assay for haplotyping by integrating allele-specific PCR [24]. This way haplotypes within PCR fragments can be measured directly.

10. QUANTITATION

The cost of SNP genotyping is still a real problem considering the number of SNPs people would like to test for whole genome association studies. Numbers of over 500.000 SNPs in 50.000 individuals have been mentioned [25]. This would mean generating 25 billion SNP genotypes for an exhaustive study of one disease. At the

current cost of SNP genotyping these studies are inconceivable. A strategy that has been suggested to alleviate this is genotyping pools of DNAs from different individuals and to apply quantitation of the results. Cases and controls would be pooled separately. The most crucial part of such a procedure is making well-defined mixtures of DNA of different individuals. We have determined that the GOOD assay can be used to identify one allele in ten individuals if pools are set up carefully [26].

Table 1: n SNPs in close proximity give rise 2^n haplotypes. With three SNPs these are 8 haplotypes with 36 possible genotypes.

SNP	1	2	3
genotype	C/G	T/A	C/G
haplotype 1	C	T	C
:			
haplotype 8	G	A	G

11. AUTOMATION OF THE GOOD ASSAY

Due to the facile processing of the GOOD assay, it is well suited for automation (Figure 8). All of the reaction steps in the preparation can be done by standard liquid handling robots.

The reaction volumes were optimised for automation of the process. The initial PCR is done in a volume of 3 µl covered with 3 µl oil (to avoid evaporation) in 384-well microtitre plates. This is at the limit of what is technically manageable by standard liquid handling robotics. The same microtitre plate is then used for successive reaction steps and the final product is transferred to a MTP™ target plate. The MTP™ target plates (Bruker Daltonik GmbH, Bremen, Germany) are in an identical layout as a standard 384-well microtitre plate. The mass spectrometer is technically capable of analysing samples at an even higher density, such as a 1536-well format.

The GOOD assay was first implemented on a RoboAmp 2000 (MWG Biotech, Ebersberg, Germany), which has an integrated PCR machine. This robot is suitable for medium-throughput applications. For high-throughput we recently implemented the GOOD assay on a BasePlate (The Automation Partnership, Royston, UK). This robot is a liquid handling robot with a 96-tip liquid handling head. Multiple 384-well microtitre plates are prepared in a single run and then manually transferred to an incubator or thermocyclers. Sample transfer from the microtitre plates to the MTP™ target plates is also done with the BasePlate.

Prepared MTP™ target plates are introduced into the MALDI mass spectrometer. Data accumulation of 384 samples is fully automatic using the Genotools software (Bruker Daltonik GmbH, Bremen, Germany). Allele calling is then done online with the SNP manager software [27].

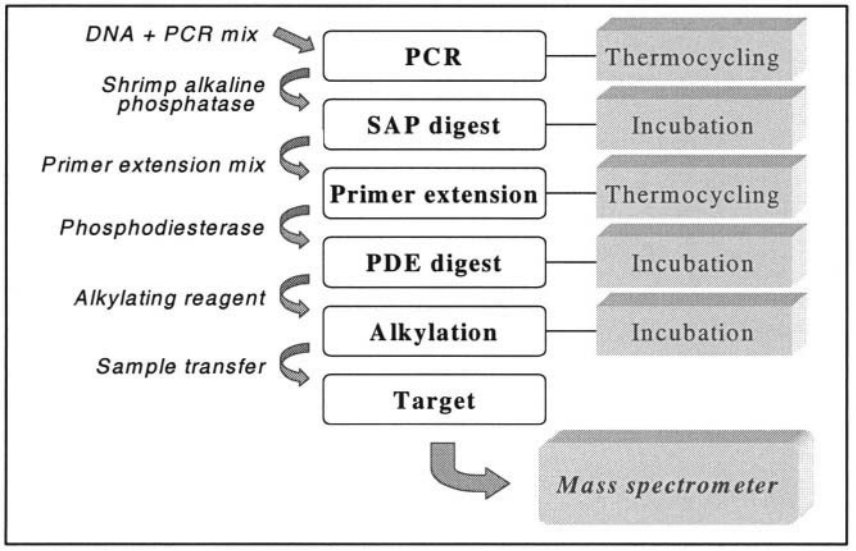


Figure 8. Reaction sequence scheme of the GOOD assay. The method consists of pipetting, incubation, and thermocycling steps and is therefore easy to automate.

Currently, commercially available MALDI mass spectrometers are capable of recording 40.000 spectra per day. The recently introduced Autoflex (Bruker Daltonik GmbH, Bremen, Germany) can be equipped with a robot that automatically changes targets. This allows 24 hour unsupervised operation. Using a conservative multiplex factor of 3, approximately 120.000 genotypes can be generated per day at a fraction of cost of other SNP genotyping technologies.

12. OUTLOOK

There are significant benefits to be gained from charge-tagging technology for the analysis of DNA, which is demonstrated by the GOOD assay. It is well adapted to the detection sensitivity of state-of-the-art MALDI mass spectrometers and the tolerances of standard automated laboratory liquid handling robotics. With automatic data accumulation and analysis systems a streamlined operation can be established. One current limitation is the low degree of multiplexing that can be provided in a single sample preparation. On the other hand data accumulation by mass spectrometry is rapid. The trade-off for establishing an experiment with a high degree of multiplexing is the time that has to be invested for the optimisation. Applications of SNP genotyping, like for the traceability of cattle, merit an effort for

the optimisation of reactions with a high degree of multiplexing, as these reaction might be used millions of times afterwards. Here it is a great benefit to have stable multiplexes of tens of SNPs.

A major drawback in the GOOD assay is the alkylating agent that is used for the chemical backbone neutralisation. Even though methyliodide is used in very low quantities, it is a hazardous reagent. We are currently investigating methods to avoid this toxic agent.

13. REFERENCES

1. Landegren U, Kaiser R, Caskey CT, Hood L. *Science* 242: 229, 1988
2. Schafer AJ, Hawkins JR *Nature Biotechnol.* 16: 33, 1998
3. Evans WE, Relling MV. *Science* 286: 487, 1999
4. Murphey LJ, Gainer JV, Vaughan DE, Brown NJ. *Circulation* 102: 829, 2000
5. Sagar M, Tybring G, Dahl ML, Bertilsson L, Seensalu R. *Gastroenterology* 119: 670
6. Gut IG, review submitted, 2000
7. Nordhoff E, Ingendoh A, Cramer R, Overberg A, Stahl B, Karas M, Hillenkamp F, Crain PF. *Rapid Commun. Mass Spectrom.* 6: 771, 1992
8. Christian NP, Colby SM, Giver L, Houston CT, Arnold RJ, Ellington AD, Reilly JP. *Rapid Commun. Mass Spectrom.* 9: 1061, 1995
9. Pieles U, Zürcher W, Schär M, Moser HE. *Nucleic Acids Res.* 21: 3191, 1993
10. Nordhoff E, Kirpekar F, Karas M, Cramer R, Hahner S, Hillenkamp F, Kristiansen K, Roepstorff P, Lezius A. *Nucleic Acids Res. 22:* 2460, 1994
11. Kirpekar F, Nordhoff E, Kristiansen K, Roepstorff P, Hahner S, Hillenkamp F. *Rapid Commun. Mass Spectrom. 9:* 525, 1995
12. Schneider K, Chait BT. *Nucleic Acids Res.* 23: 1570, 1995
13. Nordhoff E, Cramer R, Karas M, Hillenkamp F, Kirpekar F, Kristiansen K, Roepstorff P. *Nucleic Acids Res.* 21: 3347, 1993
14. Keough T, Baker TR, Dobson RL, Lacey MP, Riley TA, Hasselfield JA, Hesselberth PE. *Rapid Commun. Mass Spectrom.* 7: 195, 1993
15. Gut IG, Beck S. *Nucleic Acids Res.* 23: 1367, 1995
16. Bartlet-Jones M, Jeffery WA, Hansen HF, Pappin DJC. *Rapid Commun. Mass Spectrom.* 8: 737, 1994
17. Gut IG, Jeffery WA, Pappin DJC, Beck S. *Rapid Commun. Mass Spectrom. 11:43,* 1997
18. Berlin K, Gut IG. *Rapid Commun. Mass Spectrom.,* 13: 1739, 1999
19. Lee LG, Connell C, Woo SL, Cheng RD, McArdle BF, Fuller CW, Halloran ND, Wilson RK. *Nucleic Acids Res.* 20: 2471, 1992
20. Sauer S, Lechner D, Berlin K, Lehrach H, Escary J-L, Fox N, Gut IG. *Nucleic Acids Res.* 28: e13, 2000
21. Griffin TJ, Hall JG, Prudent JR, Smith LM. *Proc. Natl. Acad. Sci.* USA 96: 6301, 1999
22. Sauer S, Lechner D, Berlin K, Plançon C, Heuermann A, Lehrach H, Gut IG. *Nucleic Acids Res.* 28: e100, 2000
23. Lindenbaum P: http://www.cng.fr
24. Brandt O, Lechner D, Berlin K, Gut IG. *manuscript in preparation*
25. Kruglyak L. *Nature Genet.* 22: 139, 1999
26. Schatz P Diploma Thesis, Technical University Berlin, 2000
27. Pusch W, Kraeuter K-O, Froehlich T, Staalgies Y and Kostrzewa M *Biotechniques,* in press, 2001

CHAPTER 5

MICROCHIP ANALYSIS OF DNA SEQUENCE BY CONTIGUOUS STACKING OF OLIGONUCLEOTIDES AND MASS SPECTROMETRY

PH. Tsatsos, V. Vasiliskov , A. Mirzabekov
Argonne National Laboratory, Argonne, IL 60439, USA. Tel: 630-252-3981; Fax 630-252-9155; E-mail amir@everest.bim.anl.gov

1. INTRODUCTION

Microarrays of immobilised oligonucleotides are becoming a powerful instrument for analysing DNA and RNA sequences. Conner *et al.* (1983) [1] were the first to suggest using an array of a few filter-immobilised custom-synthesised oligonucleotides to analyse specific DNA mutations. The next impetus for developing large oligonucleotide microarrays was the suggestion to use sequencing by hybridisation (SBH) [2, 3, 4, 5]. In sequencing by hybridisation, DNA is hybridised with a whole set of oligonucleotides of a certain length; the overlapping of the hybridised oligonucleotides allows one to reconstruct the DNA sequence. Using generic macrochips [6] and microchips [7, 8] containing immobilised oligonucleotides may be an effective tool for SBH. The longer oligonucleotides of the generic microchip, the longer the DNA that can be sequenced. To simplify the complexity of the microchip, both Lysov *et al.,* [4] and Khrapko *et al.* [7] proposed contiguous stacking hybridisation (CSH). In this method, DNA is hybridised simultaneously, with complementary immobilised oligonucleotide of length N in the presence of a shorter oligonucleotide length n, which can extend N in its DNA duplex. The CSH produces a contiguous duplex of length N+n that contains a broken phosphodiester bond in one strand of the duplex between immobilised oligonucleotide N and a solution oligonucleotide n. Such CSH may increase the efficiency of the generic microchip containing 4^N oligonucleotide to one containing 4^{N+n} oligonucleotides. Using fluorescence to identify n contiguously stacked hybridised oligonucleotides is a rather inefficient procedure [9]. Using matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry to identify these oligonucleotides [10] radically improves the efficiency of CSH. Combined use of the generic microchip, CSH, and MALDI-TOF mass spectrometry could also provide a versatile method for proofreading and analysing mutations and single nucleotide polymorphism and sequencing by hybridisation of a large variety of DNA sequences.

J.N. Housby (ed.), Mass Spectrometry and Genomic Analysis, 66-75.

2. MAGIChip PROPERTIES

MAGIChips (Micro Arrays of Gel-Immobilised Compounds on a Chip) are arrays that have been developed over the past several years [7, 8, 11]. This type of microchip is established on a glass surface that has tiny polyacrylamide gel elements affixed to it. Each gel pad is approximately 100 μm square or round with a height of 10-50 μm. The space separating each gel element is a distance of about 200 μm (Figure 1). Each discrete gel element can function as a spot for immobilisation as well as a single test tube because the gel element is encircled by a hydrophobic glass surface that prevents the exchange of solution among the gel elements. This property permits one to perform element-specific reactions, such as ligation, phosphorylation [12], single base extension [13] and other enzymatic reactions [14] with the immobilised and hybridised substances of interest.

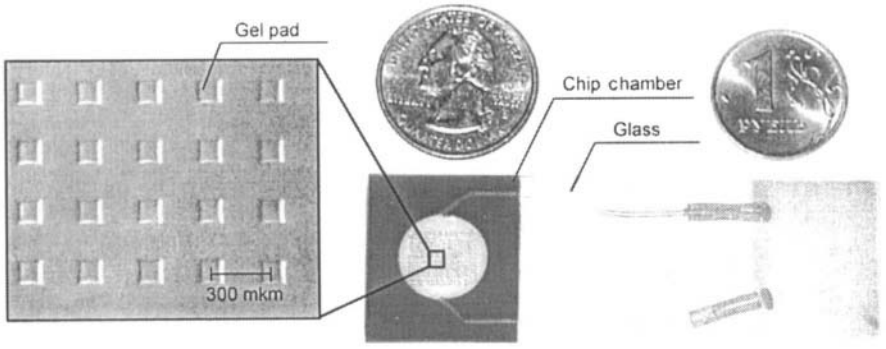


Figure 1. Diagram of MAGIChip

2.1 Production of MAGIChip

There are two alternative methods of microchip manufacturing. The first method is through intermediate production of blank microchips in which the probes are immobilised to the gel pads after photopolymerisation and activation of the acrylamide [11]. The second method involves the simpler copolymerisation procedure in which the method for the production of matrices and the step of probe immobilisation is combined [15]. In this method, acrylamide is copolymerised with oligonucleotides or proteins that contain unsaturated residues.

2.2 Activation of Probes

Probes to be immobilised in the gel pads must be activated in order to contain chemically reactive groups for coupling with the activated gel pads. For example, immobilisation in aldehyde-containing gels would require the probe to be functionalised by introduction of amino groups [16]. The probe is prepared either

by introduction of chemically active groups in terminal positions of the oligonucleotides during their chemical synthesis or within the chain of nucleotides by chemical synthesis or natural means [16, 17]. Probe activation chemistry is well developed and has both high yield and reproducibility.

2.3 Chemical Immobilisation of Probes

Two methods for immobilisation have been used in our group. In the first method, the gel supports contain amino or aldehyde groups that couple with oligonucleotides bearing aldehyde or amino groups [18]. In the second method, the polyacrylamide gel matrix is activated by introducing hydrazide groups that interact with the 3'-dialdehyde termini of activated oligonucleotides. A disadvantage of this method is that the hydrazine chemistry does not provide sufficient stability of attachment in repeated hybridisation experiments.

2.4 Preparation of the Target

Proudnikov *et al.* [17] presented a preparation method based on the introduction of aldehyde groups by partial depurination of DNA or oxidation of the 3'-terminal ribonucleoside in RNA by sodium periodate. We have routinely used this method to couple fluorescent dyes with attached hydrazine groups to the aldehyde groups. This bond is then stabilised by reduction.

3. HYBRIDISATION

3.1 Theoretical Considerations of Hybridisation

The principle underlying the use of oligonucleotides and DNA microchips is the distinction between perfect and mismatched duplexes. The effectiveness of distinction depends on many parameters, such as the length of the probe, the position of the mismatch in the probe, the AT content, and the conditions of the hybridisation reaction [19, 20]. Because of the aforementioned parameters, centrally located mismatches are easier to locate than terminal ones, and shorter probes allow for easier match/mismatch discrimination. The only drawback of using shorter probes is that overall duplex stability decreases as the length of the oligomer decreases, a phenomenon that is suspected to lead to low hybridisation signals with shorter probes.

There are significant differences in duplex stability, depending on the AT content of the analysed duplexes. These differences are a result of the difference in the stability of the AT and GC base pair (two vs. three hydrogen bonds). Stability is also sequence-dependent. Duplexes of the same AT content may have different stabilities, depending on the mutual disposition of the nucleotides. There have been many attempts to equalise the thermal stability of the duplexes of differing base compositions. These attempts include using probes of different lengths and performing the hybridisation in the presence of tetramethylammonium chloride or betaine [21].

Reaction conditions can be optimised to improve the discrimination of match/mismatch duplexes. Raising the hybridisation temperature to that of the melting temperature (T_m) will enhance the discrimination between match/mismatch duplexes but decrease hybridisation [22].

The previously mentioned differences in the stability of the matched/mismatched duplexes are valid only under equilibrium hybridisation conditions. Kinetic differences can also achieve match/mismatch distinction. Post-hybridisation washes can drastically reduce the mismatched signals without affecting the matched duplexes because of faster dissociation of mismatched duplexes.

In the work of Khrapko *et al.* [7], experimental interpretations showed that if the oligonucleotide probes are immobilised in three-dimensional gel elements, the apparent dissociation temperature is, in fact, dependent on the concentration of the immobilised oligonucleotide probes. This observation was used to derive an algorithm that allows for the normalisation of oligonucleotide matrices in which a higher concentration of AT-rich and a lower concentration of GC-rich immobilised nucleotides can be used to equalise apparent dissociation temperatures of duplexes that differ in their AT content, leading to true match/mismatch discrimination.

3.2 Hybridisation on Microchips

The hybridisation of nucleic acid on microchips is now a well-developed technology. Reliable discrimination of matched duplexes formed by 15-20 mers with both DNA and RNA from duplexes containing a single, centrally located mismatch has been achieved [23, 24]. In the above-mentioned works, the difference between fluorescent signals from two pads containing perfectly matched and mismatched oligonucleotides was much more than the experimental noise and statistical errors obtained in experimental series.

Measurements of melting curves of duplexes formed on oligonucleotide microchips were used to calculate melting temperatures and thermodynamic parameters of duplexes, including, entropy, enthalpy, and free energy of the hybridisation [25]. Experiments using the generic microchip, described in the next section, allowed for the measurement of hybridisation for matched/mismatched duplexes containing all possible mispairings in all positions of the 6-mer duplexes.

4. GENERIC MICROCHIP

The so-called generic microchip uses complete sets of small oligonucleotides (e.g., 6-mers) in order to query any target sequence. Generic microchips were originally proposed for DNA sequencing [4, 5], but inherent problems have thus far hindered their performance of such sequencing. The length of the DNA fragment that can be effectively sequenced on short oligonucleotide arrays is significantly limited because of the presence of repeats found along the DNA and because short stretches of nucleotides, like those immobilised on the microchip, can be found in many positions of a complex DNA sample. By increasing the length of the immobilised probes, one can significantly increase the length and complexity of readable DNA.

However, the increase in length of immobilised oligonucleotides by one base increases the length of the sequenced DNA by two times and the complexity of the generic oligonucleotide microchip by four times [4]. Thus, a generic microchip to read 400 nucleotide long DNA should contain 262,144 9-mers! The production of such microchips still presents a challenge.

Lysov *et al.* [4] first suggested using CSH to overcome the need for the production of such impossibly large microchips. CSH has also been suggested as a means to increase the sequencing efficiency and the length of sequenced fragments. The principle of CSH will be discussed in the following section.

5. PRINCIPLE OF CONTIGUOUS STACKING HYBRIDISATION

In CSH, the first hybridisation of the target DNA with the microchip that contains the full set of oligonucleotides of a fixed length (e.g., 8-mer) is followed by additional rounds of hybridisation with fluorescently labelled oligonucleotides of a fixed length (e.g., 5-mer) (Figure 2). These fluorescently labelled oligonucleotides will form extended duplexes between the terminal bases of the existing target DNA duplex with the immobilised probe (Figure 2). The 5-mers themselves do not form stable duplexes with DNA; however, when one or two of them are stacked to the hybridised microchip 8-mers, their base pairing with the DNA is stabilised [9]. This stabilisation is a consequence of stacking interactions between the terminal bases of both the 8-mer and the 5-mer. The 5-mers have greater sensitivity to the presence of mismatches in their duplexes than do the longer probes because of their small length.

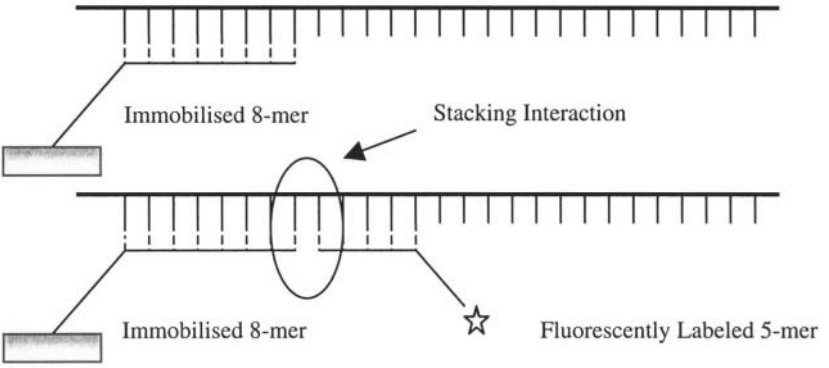


Figure 2. Principle of Contiguous Stacking Hybridisation

A full parallel screening of thermodynamic parameters of stacking between all possible interface bases X|Y was done by using our microchip, and the results were monitored using a fluorescent microscope (unpublished results). It was shown, that adenine had maximum stacking [26], but in all other cases discrimination of stacking between perfect and mismatched bases was sufficiently enough for mutation

detection. It was also shown that thermodynamic parameters of stacking on a chip closely correlated to that in solution [27].

6. MONITORING

6.1 Fluorescence

Fluorescent microscopes employing custom-designed, wide field, high-aperture, large-distance optics and a high-pressure mercury lamp as a light source for epi-ilumination were used to examine all hybridisation results obtained by using fluorescently labelled 5-mers [28]. Exchangeable filter sets allow different fluorescent labelling dyes to be used. The microscopes are equipped with a controlled temperature table allowing for temperature changes ranging from –10°C to +60°C during an experiment. A cooled CCD camera is used to record all light signals from the microchip. These light signals are then fed into the analysing computer program for quantitative evaluation of the hybridisation signals across the entire chip [25, 11]. An important advantage of the fluorescent microscope is that it allows real-time monitoring of changes in the hybridisation signal in each individual gel pad under a large variety of experimental conditions. More importantly, fluorescent microscopes allow monitoring of melting curves, which in some cases may be crucial in the proper match/mismatch discrimination.

6.2 Laser Scanner

Although the fluorescent microscope is the most widely used detection device in our experiments, it is not the only approach to microchip readout currently under development. When parallel measurements of gel pad signals are essential because of possible data loss, laser-scanning platforms are a good solution. Our scanner employs a 2mW HeNe laser as an excitation source and a low-noise PIN photodiode as a detector. The numerical aperture of the miniature objective lens is 0.62, with a working distance of approximately 3mm, which is big enough for scanning packaged microchips. All parameters of the scanning, data visualisation, and processing are set up via a host computer. We have determined that the detection threshold of the scanner is approximately two attomoles of Texas Red per gel pad, with a linear dynamic range of up to three orders of magnitude in terms of integral signal intensities.

6.3 Mass Spectrometry

A Matrix-Assisted Laser Desorption/Ionisation Time-of-Flight (MALDI-TOF) mass spectrometer was used to carry out mass spectral analysis of hybridisation results. The MALDI-TOF mass spectrometry is a rapid and accurate method for multiplex analysis of mixtures containing up to 500 short DNA molecules of different mass [29]. The effectiveness of MALDI-TOF mass spectrometry has been demonstrated for reading Sanger sequencing ladders [30] and for identifying primers with a single base [31] or by direct comparison of PCR-amplified fragments [32].

We have successfully used MALDI-TOF mass spectrometry to identify stacked 5-mers during CSH [10]. Several samples of DNA were hybridised to different microchip-immobilised oligonucleotides in the presence of a 5-mer mixture. The molecular-mass differences among any of the four nucleotides — dAp (313.2 Da), dCp (289.2 Da), dGp (329.2 Da) and dTp (304.2 Da) — constitute at least 9 Da, which is easily resolved by using MALDI-TOF.

In MALDI-TOF mass spectrometry analysis, the gel pads on the microchip must be preliminarily dried before attachment to the MALDI-TOF target plate because they are placed into the vacuum chamber of the mass spectrometer. The target with the attached microchip can be easily moved in both the x and y directions to allow each gel pad to be readily monitored. The gel elements on the microchip are mixed with a chromophoric matrix in a molar ratio of 1000:1-10000:1. (In this particular case, the matrix solution for DNA consisted of 0.5% ammonium citrate in a saturated water solution of 2-amino-5-nitropyridine.) The microchip is then inserted into the mass spectrometer and irradiated with a pulsed laser at an absorption maximum of the matrix. The interaction of photons with the matrix sample results in the formation of intact ions related to the molecular mass of the nonimmobilised DNA. As detailed in Stomakhin *et al.* [10], in order to achieve high discrimination, a CSH with unknown DNA and a mixture of all possible 5-mers was proposed where each of all possible pentanucleotides, having a known mass-label, were contiguously hybridised to an unknown sequence.

6.4 Example of Mutation Detection by CSH and MALDI-TOF Mass Spectrometry

The CSH has been successfully applied for the detection of mutations in the Rpo-B gene of Mycobacterium tuberculosis, which is responsible for rifampicine resistance in many drug resistant strains of Mycobacterium tuberculosis. A custom designed microchip was prepared. The microchip contained four gel pads with immobilised 10-mer oligonucleotides, which were complementary to four different invariable regions of the gene, so that four variable hot spots lay near the stacked ends of immobilised oligonucleotides at the region of CSH of 5-mers (Figure 3). Initially, the microchip was hybridised with DNA fragments of the gene at 30°C. Such DNA was obtained from known types of mycobacteria, amplified by nonsymmetrical PCR, and fragmented. The microchip was then hybridised with a mixture of 13 5-mers at 15°C, which is complementary to normal and all mutant DNA types and can form stacking adjacent to immobilised 10-mers. The stacking effect enlarges melting temperatures of 5-mers and allows for avoidance of cross-hybridisations at nonstacked places in other parts of DNA containing 5-mer repeats. These microchips allow for discovery of nine well-known mutations of mycobacteria at four hot spots. Another 4 of 13 remaining 5-mers are responsible for wild type DNA. After hybridisation, each microchip gel pad was incubated with matrix solution, which is needed for ionisation in MALDI-TOF MS. Hybridised 5-mers were dissociated at the surface of the gel pads into a volume of droplets, which covered each gel pad independently. The microchip was then dried and placed into MALDI-TOF MS device. As an example, Figure 3 shows only two cases, which are

responsible for two mutations and their corresponding wild types. Microchips were hybridised three times independently: the first time with wild type DNA and the last two times with two different DNAs that contained one mutation in a predetermined position. In each case a dominant peak corresponding to a wild type or mutant DNA was observed. Other minor peaks corresponding to cross hybridisation of 5-mers to other places can be predicted and removed.

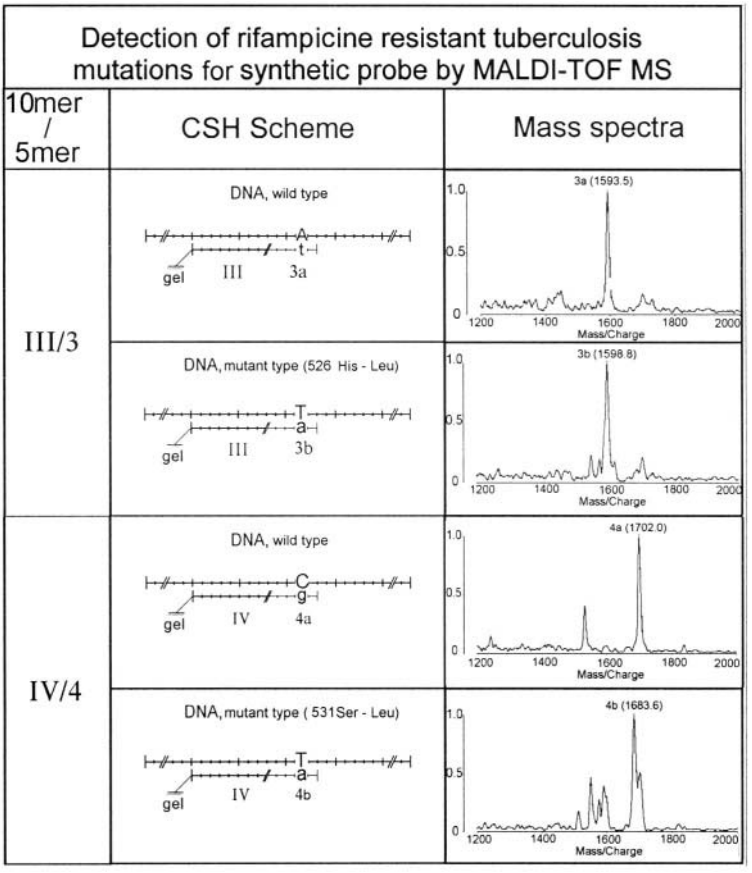


Figure 3. Mass spectra from several gel pads of CSH- MALDI-TOF MS microchip

7. CONCLUSIONS

Microchips, with their unique set of biological manipulations, could greatly enhance the analytical and diagnostic procedures currently used in biophysical laboratories. In the near future, the use of microchip applications could eliminate the need to maintain expensive complex biophysical laboratories that carry large scientific and support staff.

Gel pads of MAGIChips could serve as micro cuvettes, which would enable massive parallel or multiplex thermodynamical studies for any DNA duplexes, triplexes, DNA binding ligands, proteins, and so forth. Thousands of melting curves could be obtained simultaneously, which cannot be accomplished with the usual UV-VIS spectrophotometer. This fact demonstrates that all mismatch discrimination and mutation detection is based on the solid theoretical foundation of classical DNA thermodynamics.

Generic chip properties for long sequence detection and sequence determination in long repeats could be sufficiently expanded by the CSH approach without increasing the length of immobilised oligonucleotides (e.g., without increasing the quantity of pads on a chip). In addition, CSH uses 5-mers in variable parts of detecting DNA. Therefore, mismatch discrimination for CSH is higher than that for usual hybridisation, which uses more long oligonucleotides.

The information and research presented in this chapter effectively demonstrate a practical example of the MALDI-TOF application in the field of mutation detection. It is our hope to continue and enhance this progressive direction for microchip applications.

8. ACKNOWLEDGEMENTS

We would like to thank Andrei Stomakhin and Dennis Schulga for stimulating discussions and critical reading of this manuscript. We would also like to acknowledge the help of Felicia King and Kevin Brown for their help with editing of this manuscript. The research described here is the result of the Joint Biochip Program of Argonne National Laboratory, Argonne, IL 60439, USA, and of Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, 117984, Moscow, Russia. Panagiota H. Tsatsos is at Argonne National Laboratory. Vadim Vasiliskov is at the Engelhardt Institute of Molecular Biology and Andrei Mirzabekov is at both Argonne National Laboratory and the Engelhardt Institute of Molecular Biology.

9. REFERENCES

1. Conner BJ, Reyes AA, Morin C, Itakura K, Teplitz RL, Wallace RB. *Proc. Natl. Acad. Sci.* 8: 278, 1983
2. Bains W, Smith GC. *J. Theor. Biol*. 135: 303, 1988
3. Drmanac R, Labat I, Brukner I, Crkvenjakov R. *Genomics* 4: 114,1989
4. Lysov Y, Florentiev V, Khorlin A, Khrapko K, Shick V, Mirzabekov A. *Proc. Natl. Acad. Sci.* USSR 303: 436,1988
5. Southern EM. United Kingdom Patent Appl. GB 8810400, 1988
6. Southern EM, Maskos U, Elder JK. *Genomics* 13: 1008,1992
7. Khrapko KR, Lysov YP, Khorlin AA, Shick VV, Florentiev VL, Mirzabekov AD. *FEBS Letters* 256: 118, 1989
8. Khrapko KR, Lysov Y, Khorlin A, Ivanov I, Yershov G, Vasilenko S, Florentiev V, Mirzabekov A. *DNA Sequence* 1: 375, 1991
9. Parinov S, Barsky V, Yershov G, Kirillov E, Timofeev E, Belgovskiy A, Mirzabekov A. *Nucleic Acids Res.* 24: 2998, 1996

10. Stomakhin A, Vasiliskov V, Timofeev E, Schulga D, Cotter R, Mirzabekov A. *Nucleic Acids Res.* 28: 1193, 2000
11. Yershov G, Barsky V, Belgovsky A, Kirillov E, Kreindlin E, Ivanov I, Parinov S, Guschin D, Drobishev A, Dubiley S, Mirzabekov A. *Proc Natl Acad Sci* 93: 4913, 1996
12. Dubiley S, Kirillov E, Lysov Y, Mirzabekov A. *Nucleic Acids Res.* 25: 2259, 1997
13. LaForge KS, Shick V, Spangler R, Proudnikov D, Yuferov V, Lysov Y, Mirzabekov A, Kreek MJ. *Am. J. Med. Genet.* (Neuropsychiatr. Genet.) 96: 604,2000
14. Arenkov P, Kukhtin A, Gemmell A, Voloshchuk S, Chupeeva V, Mirzabekov A. *Anal. Biochem.* 278: 123, 2000
15. Vasiliskov AV, Timofeev EN, Surzhikov SA, Drobyshev AL, Shick VV, Mirzabekov AD *Biotechniques* 27: 592, 1999
16. Proudnikov D, Timofeev E, Mirzabekov A. *Anal. Biochem.* 259: 34, 1998
17. Proudnikov D, Mirzabekov A. *Nucleic Acids Res.* 24: 4535, 1996
18. Timofeev E, Kochetkova S, Mirzabekov A, Florentiev V. *Nucleic Acids Res.* 24: 3142, 1996
19. Livshits M, Florentiev V, Mirzabekov A. *J. Biomol. Struct. Dyn.* 11: 783, 1994
20. Livshits M, Mirzabekov A. *Biophys. J.* 71:2795, 1996
21. Mirzabekov A. *Trends Biotech.* 12: 27, 1994
22. Zlatanova J, Mirzabekov A. *Methods in Molecular Biology. Vol 170: DNA Arrays: Methods and Protocols.* Totowa, N.J.: Humana Press Inc., in press. 2000
23. Bavykin SG, Akowski JP, Zakhariev VM, Barsky VE, Mirzabekov AD. in press. 2000
24. Drobyshev A, Mologina N, Shick V, Pobedimskaya D, Yershov G, Mirzabekov AD. *Gene* 188: 45, 1997
25. Fotin A, Drobyshev A, Proudnikov D, Perov A, Mirzabekov A. *Nucleic Acids Res.* 26: 1515, 1998
26. Bommartio S, Peyret N, SantaLucia Jr. J. *Nucleic Acids Res.* 28 : 1929, 2000
27. Lysov Y, Chernyi A, Balaeff A, Beattie K, Mirzabekov A. *J. Biomol. Struct. Dyn.* 11: 797, 1994
28. Barsky I, Grammatin A, Ivanov A, Kreindlin E, Kotova E, Barskii V, Mirzabekov A. *J. Opt. Technol.* 65: 938, 1998
29. Graber JH, Smith CL, Cantor CR. *Genet. Anal.* 14: 215, 1999
30. Fu D-J, Tang K, Braun A, Reuter D, Darnhofer-Demar B, Little DP, O'Donnell MJ, Cantor CR, Koster H. *Nature Biotechnol.* 16: 381 1998
31. Fei Z, Ono T, Smith LM. *Nucleic Acids Res.* 25: 2827, 1998
32. Haff LA, Smirnov IP. *Nucleic Acids Res.* 24: 3749, 1997

CHAPTER 6

SHORT OLIGONUCLEOTIDE MASS ANALYSIS (SOMA)

an esi-ms application for genotyping and mutation analysis

PE. Jackson, MD. Friesen, JD. Groopman
Johns Hopkins University, Department of Environmental Health Sciences, 615 N. Wolfe Street, Baltimore, MD 21205, USA. Tel: 410-955-4235; Fax: 410-955-0617; E-mail pjackson@jhsph.edu and Unit of Nutrition and Cancer, NTR International Agency for Research on Cancer, 150 cours Albert Thomas F-69372 LYON Cedex08, FRANCE. E-mail friesen@iarc.fr

1. INTRODUCTION

We have developed a method for analysis of genetic alterations in PCR amplified DNA fragments using electrospray ionisation mass spectrometry (ESI-MS) termed Short Oligonucleotide Mass Analysis (SOMA). The technique is highly accurate and has been applied to genotyping individuals and to mutational analysis of human tumours.

Several methodologies for analysing DNA variants by mass spectrometry (MS) have been developed. The length of the DNA fragments in which a single base change can be measured is limited by the mass resolution of the instrument. Therefore, in order to detect the smallest mass difference occurring in a genetic alteration (9 Da between adenine and thymine), methods are required which generate short oligonucleotides (<20 bases) for analysis. Several groups have developed methods in which oligonucleotides are hybridised to a PCR product template and then analysed by MS [1-7]. An alternative approach is to use restriction endonucleases to digest PCR products and analyse the resultant fragments directly by MS [8-10]. Such a method whereby short, defined fragments are produced by digestion using a type IIS restriction endonuclease and analysed by ESI-MS is described in this chapter.

2. SHORT OLIGONUCLEOTIDE MASS ANALYSIS

2.1. Method Outline

SOMA employs PCR amplification using primers that contain the recognition sequence for a type IIS restriction endonuclease, which is incorporated into the

J.N. Housby (ed.), Mass Spectrometry and Genomic Analysis, 76-92.

resultant PCR product on either side of the site of interest (Figure 1). Type IIS restriction endonucleases, such as *BpmI*, have a recognition sequence distal to their cleavage site, enabling mismatches to be designed in the PCR primers at a suitable distance from the 3′ end of the primer such that annealing is not affected and the cleavage site is close to the site of interest. Following amplification, the PCR products are digested with the relevant restriction enzyme to release an oligonucleotide containing the site to be interrogated and the sample is purified using a simple phenol-chloroform extraction and ethanol precipitation.

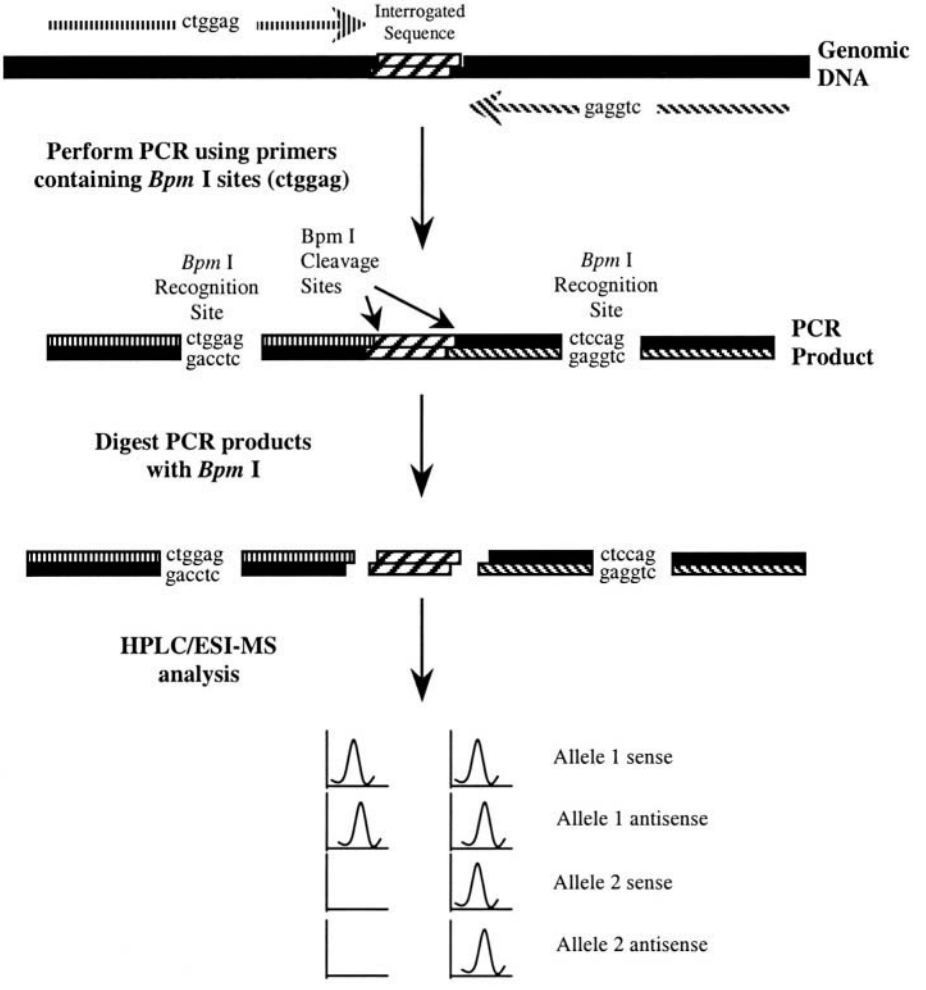


Figure 1. Strategy for the preparation of DNA suitable for SOMA. (Adapted from [10])

The oligonucleotides are then introduced into the ESI-MS using an HPLC, which allows further in-line clean up of the sample, for analysis. Ion masses representing both the sense and antisense strands for each variant are monitored independently by ESI-MS, providing a near-perfect internal control.

2.2. Design of PCR Primers and Fragments for Analysis

The primers used for PCR amplification are designed to flank the site of interest and contain the recognition sequence for a type IIS restriction enzyme. The enzyme we use is *BpmI*, (New England Biolabs, Beverly, MA, USA) which has a recognition sequence of CTGGAG and a cleavage site 16 and 14 bases upstream on the upper and lower strands, respectively. The recognition sequence for the enzyme is therefore placed 16 bases from the end of the primer, replacing the bases normally present at this position in the genome. As mismatches with the DNA template may result from the creation of the recognition sequence, the primers used are 35-40 bases long to ensure high fidelity annealing to the flanking sequence. In addition, as all PCR primer pairs tested give robust products using the same cycling temperatures, optimisation of reaction conditions is not generally required.

Two factors must be taken into account when positioning the enzyme recognition sequence around the variant sequence under study: (a) length of the DNA fragment to be generated following digestion and (b) the position of the variant sequence within the DNA fragment.

Oligonucleotides ranging from 7 to 17 bases have been successfully ionised and resolved using ESI-MS, however, for optimum mass resolution, it is preferable to keep oligonucleotides small. The charge state of an ion tends to increase with mass and can depend on the instrument used. For example, on the Quattro LC (Micromass, Manchester, UK), 7- 8- and 9-mer oligonucleotides all tend to give predominantly doubly charged ions. A 10-mer oligonucleotide, however, gives a mixture of doubly and triply charged ions, leading to a decrease in sensitivity if only one of the two charge states is monitored. Multiplex SOMA can be carried out by designing primers for each of the multiple variant sequences such that oligonucleotides of differing length are generated for each variant. In this way, mass spectral peaks are well distributed along the mass scale and overlapping of peaks is avoided. DNA fragments should also be designed to avoid generating fragments with the same molecular weight and the formation of metal-ion adducts on oligonucleotides should be considered, e.g., would a sodium adduct on the sense strand of one allele increase its mass such that it fell in the window monitored for one of the other oligonucleotides?

DNA fragments should be designed such that the variant sequence under study is positioned at or near the 5' end of the sense or antisense strand. In this way, intense, sequence specific MS/MS fragment ions are generated by collision-induced dissociation. Positioning the variant sequence near the middle of a DNA fragment

may result in relatively weak sequence-specific fragment ions and intense non-sequence specific ions.

2.3. Typical PCR Reaction Conditions

The same PCR reaction conditions have been used to generate robust products for a majority of primer pairs and templates examined to date. This has been possible because the PCR primers used were long and the fragments amplified were only 80-l00bp in length,. The 50μl PCR reaction mix contained 16.6mM NH_4SO_4, 67mM Tris-HCl (pH 8.8), 6.7mM $MgCl_2$, 10mM β-mercaptoethanol, 4μl DMSO, all four deoxynucleoside triphosphates (each at 0.6mM), 350ng of each primer and 0.5 units Platinum Taq polymerase (Life Technologies (Gibco BRL), Rockville MD, USA). Reactions were performed with 25-50ng genomic DNA or 10μl of DNA eluate from plasma samples. Thermocycling conditions were 95°C for 2 mins, followed by 40 cycles of 95°C for 30s, 65°C for 30s and 72°C for 30s.

3. ELECTROSPRAY IONISATION MASS SPECTROMETRY

3.1. Formation of ions

Elecrospray ionisation mass spectrometry is one of two methods used for ionising and analysing large biomolecules such as DNA and proteins, the other being matrix-assisted laser desorption ionisation (MALDI) [11; 12]. During ESI-MS, the analyte of interest is introduced into the mass spectrometer as a solution. A high voltage is applied to the narrow capillary through which the sample is introduced into the MS, resulting in the formation of charged droplets at the end of the needle. As the solvent evaporates, electrostatic repulsion causes the droplets to break apart into individually charged molecules with one or multiple charges [13]. The ions are accelerated by electric fields towards the mass analyser, typically a quadrupole or ion trap.

3.2. Tandem mass spectrometry

Tandem mass spectrometry, or MS/MS, provides an additional level of specificity for analysing compounds and is particularly useful for large biomolecules such as oligonucleotides and proteins. The parent ion of interest is isolated and subject to collision with an inert gas (such as helium or argon), causing the ion to fragment into sequence specific daughter ions. One or several of the resultant daughter ions can be monitored with time. If other compounds have the same mass to charge (m/z) ratio as the parent ion they will also be isolated during the first stage, but will fragment into daughter ions with different m/z ratios and will therefore not be detected. Thus, appropriate selection of the daughter ions monitored enables the distinction between two compounds with the same m/z ratio, such as two oligonucleotides with the same base composition but different sequences. MS/MS can be performed on triple quadrupole or ion trap mass spectrometers.

3.3. Typical ESI-MS Settings for SOMA

The mass spectrometer used for the initial development of SOMA and to obtain most of the results presented here was a Finnigan LQC ion trap (Thermoquest, San Jose, CA, USA) equipped with an ESI source. SOMA has also been successfully implemented using a Quattro LC, a triple quadrupole ESI-MS (Micromass, Manchester, UK).

ESI-MS of oligonucleotides was performed in the negative ionisation mode using hexafluoro-2-propanol and methanol as the mobile phase. Typical settings for the LCQ are a spray voltage of –2.5 to –5kV, a heated capillary temperature of 150 to 180°C, a capillary voltage of –40 to –60V and a sheath gas flow rate of 50 to 70 arbitrary units. The ion optics were typically 5 to 7V for the octapole 1 offset, 15 to 40V for the lens voltage and 7 to 9V for the octapole 2 offset.

4. PURIFICATION PROCEDURES

A major consideration when using mass spectrometry for analysis of DNA is the formation of Na and K adduct ions, which reduce both resolution and sensitivity. It is therefore very important to reduce the prevalence of these adduct ions during sample preparation [reviewed in ref. 14]. As PCR amplification and restriction digestion require buffered solutions containing cations, it is necessary to purify the samples prior to their introduction into the mass spectrometer.

4.1. Phenol/Chloroform Extraction and Ethanol Precipitation

An easy clean-up method that has been used by several groups for preparation of samples for MS analysis is phenol/chloroform extraction followed by ethanol precipitation. We have found this to be the most effective method with good sample recovery and low background contamination. The use of ammonium acetate during ethanol precipitation results in an exchange of sodium ions with ammonium ions, which are not observed to form adducts with oligonucleotides [15; 16].

During purification, the volume of the digest reaction is increased to 100µl with water and an equal volume of phenol/chloroform added. After mixing thoroughly, the samples are centrifuged and the aqueous phase removed to a fresh tube. DNA is then precipitated in the presence of 3µl SeeDNA (Amersham, Piscataway, NJ, USA), one-tenth volume of 7.5M ammonium acetate and 3 volumes of ethanol. After mixing, the DNA is pelleted by centrifugation and the pellet washed with 70% ethanol. For maximal sensitivity, the DNA pellet can be resuspended in water and a second precipitation performed prior to washing with 70% ethanol. Using synthetic oligonucleotides to spike reactions before and after purification, the recovery of oligonucleotide was found to be 98%.

4.2. In-line HPLC Purification

Introducing the sample into the ESI-MS with an HPLC provides an in-line method of purification. Initially, the oligonucleotides are concentrated at the head of the column with a low organic solvent concentration while low molecular weight impurities are washed through. Oligonucleotides of interest are eluted with an increasing gradient of organic solvent. High molecular weight proteins and oligonucleotides from the enzyme digestion can be diverted from the MS ion source to reduce contamination. In this way we have been able to further purify the PCR generated samples following phenol/chloroform extraction and ethanol precipitation.

Another major advantage of using HPLC is that solvent additives can be used which help to suppress the formation of cation adducts. Apffel and co-workers reported the use of hexafluoro-2-propanol (HFIP) for ESI-MS analysis of oligonucleotides up to 74 bases long [17]. Addition of HFIP to a water/methanol gradient allows good HPLC separation of oligonucleotides and efficient electrospray ionisation with a minimum of cation adducts formed.

HPLC solvents were prepared from a stock solution of aqueous 0.8M HFIP, pH 6.8 (adjusted with triethylamine), diluted to 0.4M with water for solvent A and methanol for solvent B [10; 17]. The gradient generally used for SOMA analysis was 60% A:40% B programmed to 40% A:60% B in 5 mins, returning to the initial conditions in 0.5 mins.

5. GENOTYPING USING SOMA

The human genome contains many polymorphic sequences, i.e. sequences that differ from one individual to another. While most of these variants are not thought to have biological consequences, several have been linked to disease, such as variants in the BRCA genes that predispose to breast cancer [18] and a polymorphism in prothrombin predisposing to a bleeding disorder [19]. Testing for the presence of such polymorphisms can provide critical diagnostic information for management of patients and their families.

Considerable effort has been made in developing suitable methodologies for genotyping individuals in a clinical setting, in which major considerations are accuracy, cost and throughput. Techniques developed using mass spectrometry have received attention owing to the advantages they offer, including high accuracy, throughput and potential for automation [20; 21].

5.1. APC Genotyping in Human Subjects

We applied SOMA to genotyping several variant sites in the human adenomatous polyposis coli (*APC)* gene. The first of these is a variant (I1307K) present in 6% of the Ashkenazi Jewish population that is associated with an increased risk of colorectal cancer of approximately two fold [22]. The polymorphism is an A to T change at codon 1307 and hence represents the most difficult alteration to analyse because there is only a 9Da mass difference between the two variants. PCR primers were designed to yield 15-mer oligonucleotides for MS analysis following

restriction digestion [10]. The triply charged ion for the wild type sense (ATA-s: m/z 1578.7) and antisense (ATA-as: m/z 1522.3) oligonucleotides and the mutant sense (AAA-s: m/z 1581.7) and antisense (AAA-as: m/z 1519.3) oligonucleotides were monitored. Examples of chromatograms from two individuals are shown in Figure 2.

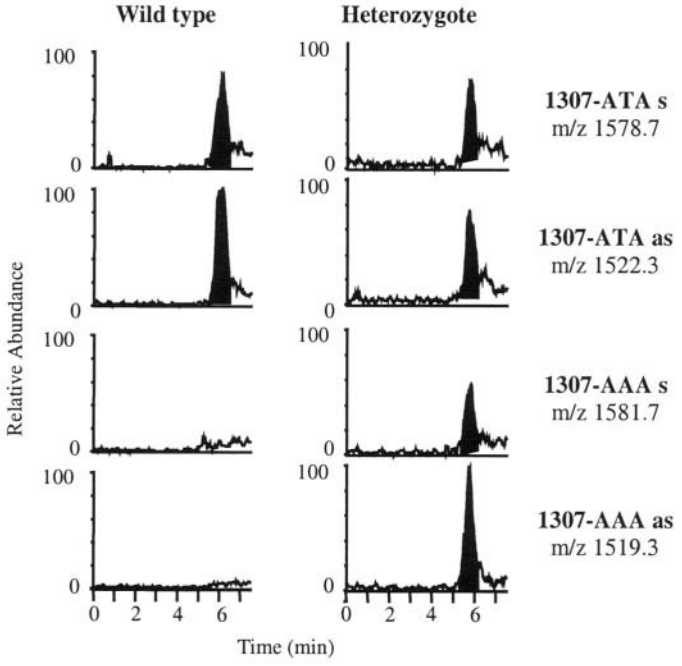


Figure 2. ESI-MS analysis of the APC codon 1307 variants. (Adapted from ref. 10)

Samples with wild type sequence only had a signal in the channels monitoring masses representative of both wild type sense (ATA-s) and wild type antisense (ATA-as) alleles. A sample for an individual who is heterozygous at this locus has signals in the sense and antisense channels for both the wild type (ATA) and mutant (AAA) alleles. In a blind analysis of 16 individuals by SOMA and sequencing there was 100% concordance.

The current test for the I1307K polymorphism uses allele-specific oligonucleotide hybridisation in which labelled probes are used to discriminate between wild type and mutant sequences. Such differential hybridisation requires considerable optimisation and is rarely perfect, resulting in a background signal that

reduces specificity and accuracy. In contrast, the results obtained using SOMA are easy to interpret and highly accurate.

A second variant in the human *APC* gene (ACA or ACG at codon 1493) was examined by SOMA to demonstrate the method could be easily applied to other polymorphic sequences. In this instance, the sense and antisense oligonucleotides for one of the alleles had the same mass, but different sequences [10]. MS/MS was therefore used for this analysis and suitable daughter ions were selected to enable distinction between the two oligonucleotides (Figure 3). Again, a blinded analysis of 50 individuals for polymorphisms at this locus was compared to sequence data and was in complete agreement.

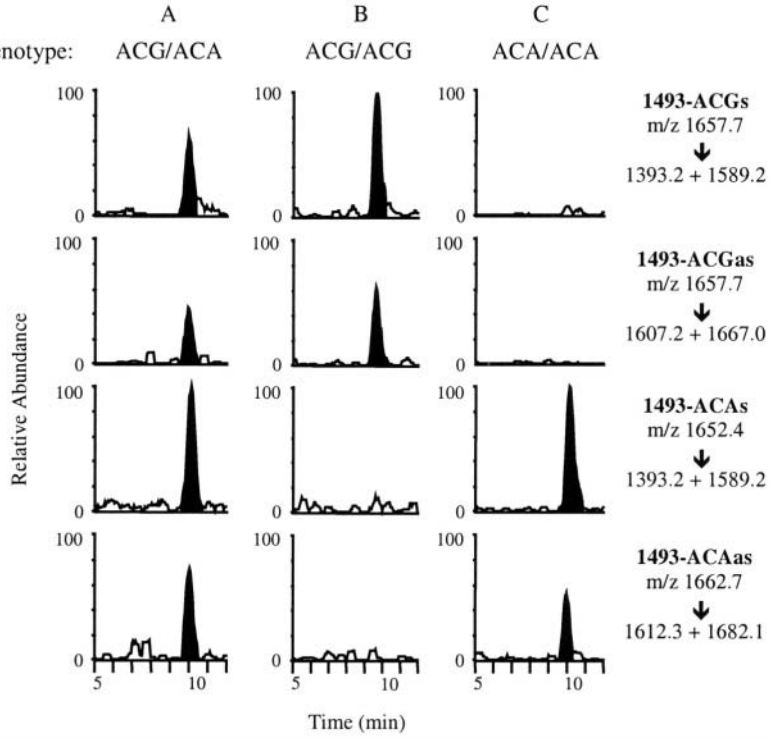


Figure 3. ESI-MS/MS analysis of the APC codon 1493 variant. (Adapted from [10])

As a preliminary examination of the potential for multiplex genotyping by SOMA, three common polymorphisms in the *APC* gene at codons 485, 545 and 1756 were analysed in parallel [10]. The PCR amplification and digest reactions were performed individually to yield fragments of 8, 9 and 11 bases for codons 485, 545 and 1756, respectively. All the reaction products for one individual were then

combined, purified by phenol/chloroform extraction and ethanol precipitation and analysed by HPLC/ESI-MS. Twelve masses representing the doubly charged ion for each of the sense and antisense alleles were monitored by ESI-MS. Simultaneous determination of polymorphisms at the three codons was possible and results for two individuals are shown in Figure 4. Results obtained by SOMA and DNA sequencing were, again, perfectly concordant.

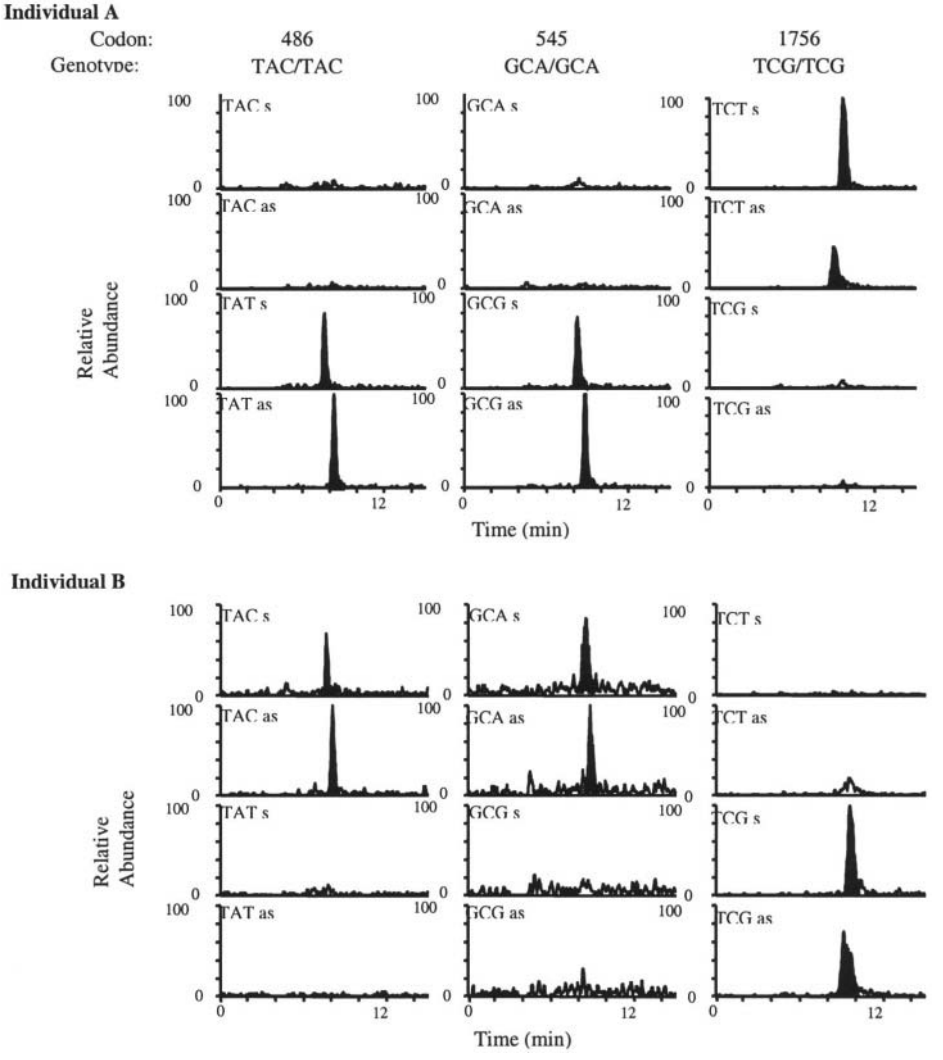


Figure 4. Simultaneous analysis of three different APC variants for two patients. (Adapted from[10])

5.2. APC Genotyping in Min Mice

Min mice are used as an animal model of familial adenomatous polyposis (FAP), an inherited disorder in which patients are highly susceptible to colorectal cancer. FAP patients have germline mutations in the *APC* gene, the gatekeeper of colorectal cancer, which leads to development of hundreds of benign colorectal polyps in these individuals [23]. FAP patients are almost certain to develop colorectal cancer due to the large number of precancerous polyps that may progress to invasive lesions.

Min mice have a truncating mutation in the murine *APC* gene at a position similar to that found in many FAP patients and develop multiple intestinal adenomas [24]. Heterozygous $APC^{min/+}$ mice were therefore used to evaluate the efficacy of potential chemopreventative compounds on development of intestinal polyps [25]. The compounds tested were sulindac, a non-steroidal anti-inflammatory drug with established chemopreventative activity, and EKB-569, an irreversible inhibitor of the epidermal growth factor receptor kinase.

The mice were genotyped using SOMA to identify heterozygous mice carrying the mutation (TTA) from those with wild type *APC* sequence (TTT) for inclusion in the study. Initial experiments to genotype using allele specific PCR were found difficult to optimise and were not robust. In contrast, SOMA was easy to implement and robust signals were obtained with the first PCR primers designed. The length of the fragments analysed was 8 nucleotides and the sequences for each fragment are shown in Table 1.

Table 1. Oligonucleotides used for genotyping min mice

Allele	Sequence	Molecular Mass	M/z of ion monitored	Daughter ions monitored
Min-ttt-s	5′-AGTTTGGA-3′	2544.6	1271.3	1030.0 + 1418.2 + 1731.2
Min-ttt-as	5′-CAAACTTC-3′	2433.6	1215.8	914.4 + 1092.3 + 1405.3
Min-tta-s	5′-AGTTAGGA-3′	2553.6	1275.8	1034.3 + 1427.1 + 1741.0
Min-tta-as	5′-CTAACTTC-3′	2424.6	1211.3	914.3 + 1083.3 + 1396.2

The doubly charged parent ion for each oligonucleotide was isolated in turn and subject to a collision energy of 30%. All the daughter ions with an m/z between 850 and 1800 were monitored in full scan mode. The sum of three specific ions for each allele was displayed to determine the genotype of the mouse and examples of the mass chromatograms obtained are shown in Figure 5.

Polyp formation was significantly reduced with sulindac alone (~70% reduction with a dose of 20mg/kg) and EKB-569 (87% reduction with a dose of 20mg/kg). Moreover, when a lower dose of sulindac was administered in combination with EKB-569, >95% reduction in polyp number was observed [25]. These results

indicate a new strategy for the chemoprevention of colorectal neoplasm's that may be applicable to humans.

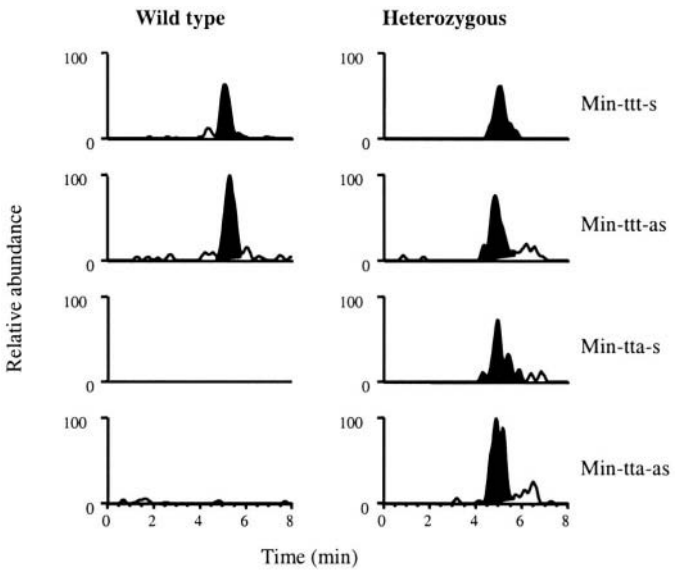


Figure 5. ESI-MS/MS genotyping of min mice.

6. MUTATION DETECTION USING SOMA

6.1. Analysis of p53 Mutations in Liver Cancer Patients

Hepatocellular cancer (HCC) is a common cause of cancer death worldwide with a striking geographical variation in incidence. In the People's Republic of China, HCC has an incidence rate of 150 cases per 100,000 per year in some regions and accounts for over 250,000 deaths annually. Major etiological factors associated with development of HCC include infection with hepatitis B or C viruses and exposure to dietary aflatoxin B_1 (AFB_1) [26; 27].

A specific mutation in the *p53* tumour suppressor gene has been detected in 10-70% of HCCs from areas of high AFB_1 exposure and is absent from HCCs from areas with negligible AFB_1 exposure [28-30]. Support for the implication that the mutation, a G→T transversion at the third base of codon 249, is caused by exposure to aflatoxin has come from in vitro studies. Aflatoxin exposure in bacteria almost exclusively causes G→T transversions [31] and the aflatoxin-epoxide has been shown to bind to codon 249 of *p53* in vitro [32]. Moreover, human

hepatocarcinoma cells exposed to aflatoxin in the presence of rat liver microsomes had a high prevalence of G→T transversions in codon 249 of the *p53* gene [33; 34]. We considered this aflatoxin-specific *p53* mutation an ideal somatic mutation to which SOMA could be applied as we have an interest in developing markers of HCC for use in intervention studies and early detection of disease.

6.1.1. p53 Mutations in Liver Tumours

In an initial validation of SOMA for analysing the codon 249 mutation in *p53*, we analysed a series of 25 paired HCC tumour and normal samples that have previously been sequenced for mutations in the *p53* gene [35]. The oligonucleotides generated following amplification and restriction digestion for ESI-MS analysis are shown in Table 2.

Table 2. Oligonucleotides used for p53 codon 249 mutation analysis

Allele	Sequence	Molecular Mass	m/z of ion monitored	m/z of daughter ions monitored
AGG-s	5′-CGGAGGCC-3′	2515.6	1256.8	1047.3 + 1181.7 + 1566.0
AGG-as	5′-CCTCCGGT-3′	2441.5	1219.8	1268.6 + 1347.8 + 1637.2
AGT-s	5′-CGGAGTCC-3′	2490.6	1244.3	899.2 + 1437.4 + 1542.4
AGT-as	5′-ACTCCGGT-3′	2465.6	1231.8	1075.0
AGA-s	5′-CGGAGACC-3′	2498.4	1248.8	1404.0 + 1693.1
AGA-as	5′-TCTCCGGT-3′	2455.4	1227.3	979.3 + 1286.0 + 1652.2

For these oligonucleotides, the doubly charged parent ions were isolated sequentially and subjected to 30% collision energy. The signals from one to three specific daughter ions were summed to provide a robust signal with a low background.

A specific G→T mutation was detected in 10/25 (40%) of the tumour samples and 0/25 of the adjacent, histopathologically normal liver tissue samples analysed by SOMA [35]. The same 10 samples were found to contain a G→T mutation by DNA sequencing [36]. However, readable DNA sequence was not generated for one of the samples found to be negative by SOMA. No G→A mutations were found in any of the samples, either by DNA sequencing or SOMA. Thus, SOMA was found to have greater sensitivity compared with DNA sequencing for the measurement of codon 249 *p53* mutations as data were obtained for all 25 patients.

6.1.2. p53 Mutations in Plasma Samples

A series of 20 liver tumour samples and plasma pairs from the same patients were obtained during a study carried out in Qidong, People's Republic of China. DNA was extracted from both tumour tissue and plasma and analysed for G→T mutations in the *p53* gene using SOMA [35]. The oligonucleotides shown in Table 2 were used, with the exception that the AGA mutation was not analysed. Representative mass chromatograms for samples with and without a mutation are shown it Figure 6.

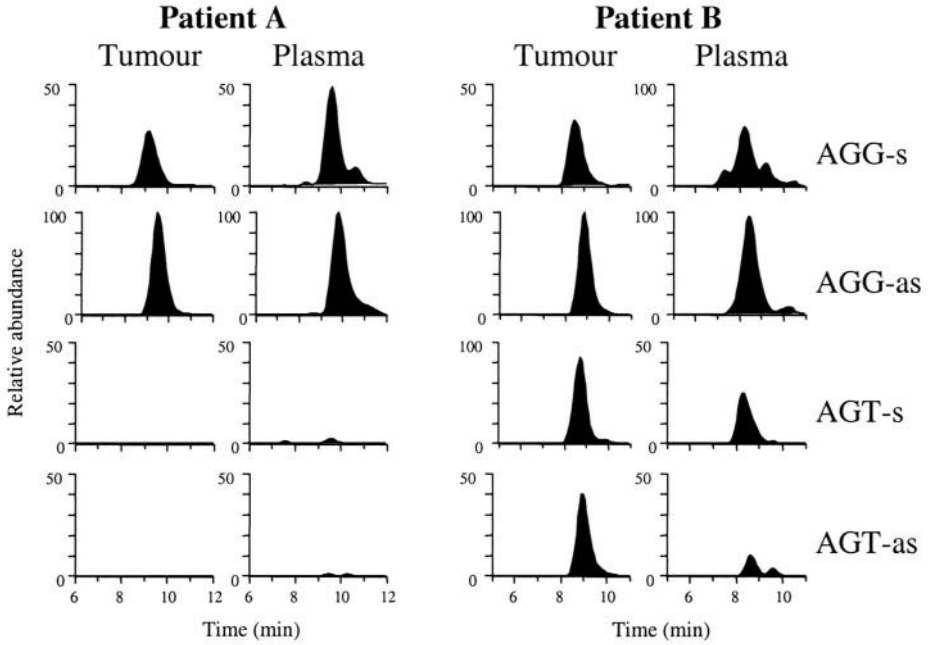


Figure 6. ESI-MS/MS analysis of p53 codon 249 mutations in paired tumour and plasma samples (Adapted from [35])

The prevalence of G→T mutations in the tumour samples was 55% (11/20), similar to that found in a previous study of patients from Qidong [36]. Paired plasma samples were also analysed and a summary of the results obtained is shown in Table 3.

A G→T mutation was found in 55% (6/11) of plasma samples in patients whose liver tumour tissue was also positive, indicating for the first time a relation between plasma and tumour pairs for the occurrence of specific *p53* mutations. An earlier study by Kirk *et al* [37] reported the detection of codon 249 *p53* mutations in the

plasma of liver tumour patients from The Gambia using a PCR-restriction endonuclease gel-based method. However, the mutational status of the tumours was not known so a direct comparison between plasma and tumour samples could not be made.

Table 3. p53 codon 249 G→T mutations in paired plasma and tumour samples

	Positive plasma	Negative plasma
Positive tumour	6 (30%)	5 (25%)
Negative tumour	4 (20%)	5 (25%)

Interestingly, four plasma samples had a detectable *p53* mutation while none were detected in the paired tumour sample. This is most likely to be indicative of multiple independent HCC nodules in these patients, one of which has a *p53* mutation. An additional ten plasma samples obtained from healthy US controls were also analysed for the specific *p53* mutation. For these samples, a larger amount of plasma (300μl vs. 100μl for the Chinese samples) was required for DNA extraction in order to obtain a robust wild type signal. No mutations were detected in any of these samples indicating false positives are unlikely to occur using SOMA. The absence of a mutation in five of the plasma samples from patients with a positive tumour sample may be due to insufficient sensitivity and future work will determine the practical limit of detection of the assay.

7. ADVANTAGES AND DISADVANTAGES OF SOMA

SOMA is a mass spectrometry-based method that can be used for analysing known variants, such as polymorphisms or mutations, and offers several advantages over traditional gel-based methodologies. Accuracy and specificity are considerably increased because an intrinsic property of the PCR product, i.e. molecular mass, is analysed directly, rather than relying on hybridisation or mobility through a gel. In addition, both sense and antisense strands are analysed in parallel, providing additional reliability, and there is no use of radiolabels or fluorescent dyes. The assay can be designed to analyse any subtle variation of interest; even the smallest possible mass difference (i.e. between A and T) can be measured with excellent resolution, and, importantly, there is no limitation posed by a requirement for existing restriction sites to be present near the sequence of interest. The results generated are accurate, sensitive and very easy to interpret.

SOMA is not appropriate for high-throughput screens to identify variants that may occur in numerous positions within a gene as other technologies exist which are better able to address such questions, for example chip [38] or gel-based technologies. Rather, SOMA is more suited to applications in which accuracy and reproducibility are critical factors.

A consideration to bear in mind when using mass spectrometry for DNA analysis, either by SOMA or other methods, is that the presence of Na and K can result in the formation of adduct ions. The presence of adduct ions reduces both sensitivity and resolution and purification steps are therefore required to minimise their occurrence during sample preparation. Another concern of mass spectrometry-based methods is cost. However, once the mass spectrometer has been purchased (~US$ 100,000 to 300,000), the cost per sample is similar to that for other technologies.

8. FUTURE PERSPECTIVES

Up until now, the SOMA method has been used only in applications such as genotyping or variant DNA detection, where PCR-based approaches are available and new high-throughput techniques are under development or being used. Except for applications that require the highest level of accuracy, SOMA will probably not be able to compete with these techniques for these applications.

However, the signal measured by SOMA is quantitative in nature as a given amount of sample is introduced into the mass spectrometer. Sample analysis by MALDI-MS requires ionising an area of the sample spot using a laser beam and there is considerable heterogeneity in sample concentration within the spot. While advances have been made to improve spotting techniques and reproducibility for MALDI [39], ESI-MS sample introduction is currently more homogeneous and therefore amenable to quantitation. We envision developing SOMA for quantitation of low levels of mutations in a background of wild type alleles. One area of particular interest is to evaluate the potential use of tumour suppressor gene and oncogene mutations as efficacy markers in chemoprevention trials or as early detection methods for cancer development.

Preliminary data on using SOMA for determining allele frequency within a population is encouraging (Friesen et al, unpublished). DNA from 100-500 individuals is pooled and the relative frequency of polymorphic alleles can be determined from the size of the respective mass chromatogram peaks. These data provide valuable information for epidemiological studies on the effect of certain polymorphisms on disease risk.

The highly accurate and quantitative aspects of SOMA are qualities suited for certain clinical diagnostic applications. For example, the quantity of a given mutant allele in a sample may determine the intervention or therapeutic strategy adopted for that individual. With continued improvements in instrumentation, data analysis capabilities and sample preparation methods, SOMA has great potential as a method for analysis of genetic variations in research and clinical settings.

9. ACKNOWLEDGEMENTS

Financial support for this work was provided by the Clayton Fund, grants P01ES06052, NIEHS Centre P30 ES03819, NCI Centre P30 CA06973, NCI grants CA43460, CA57345, CA62924, and Association for International Cancer Research grant 94275 and support from Thermoquest Corporation.

10. REFERENCES

1. Braun A, Little DP, Koster H. *Clin Chem.* 43: 1151, 1997
2. Braun A, Little DP, Reuter D, Muller-Mysok B, Koster H. *Genomics.* 46: 18, 1997
3. Little DP, Braun, A, Darnhofer-Demar, B, Koster H. *Eur J Clin Chem Clin Biochem.* 35: 545, 1997
4. Haff LA, Smirnov IP. *Genome Res.* 7: 378, 1997
5. Ross P, Hall L, Smirnov I, Haff L. *Nat Biotechnol.* 16: 1347, 1998
6. Griffin TJ, Tang W, Smith LM. *Nat Biotechnol.* 15: 1368, 1997
7. Griffin TJ, Hall JG, Prudent JR, Smith LM. *Proc Natl Acad Sci. U S A,* 96: 6301, 1999
8. Naito Y, Ishikawa K, Koga Y, Tsuneyoshi T, Terunuma H, Arakawa R. *Rapid Commun Mass Spectrom,* 9: 1484, 1995
9. Tsuneyoshi T, Ishikawa K, Koga Y, Naito Y, Baba S, Terunuma H, Arakawa R, Prockop DJ. *Rapid Commun Mass Spectrom.* 11: 719, 1997
10. Laken SJ, Jackson PE, Kinzler KW, Vogelstein B, Strickland PT, Groopman JD, Friesen MD. *Nat Biotechnol.* 16: 1352, 1998
11. Busch KL. *J Mass Spectrom.* 30: 233, 1995
12. Stults JT. *Curr Opin Struct.Biol.* 5: 691, 1995
13. Kebarle P, Yeunghaw H. On the mechanism of electrospray mass spectrometry. In: R.B. Cole (ed.), Electrospray ionization mass spectrometry. Fundamentals, instrumentation and applications, pp. 3-63, New York: *John Wiley & Sons,* Inc. 1997
14. Guo B. *Anal Chem.* 71: 333R, 1999
15. Stults JT, Marsters JC. *Rapid Commun Mass Spectrom.* 5: 359, 1991
16. Potier N, Van Dorsselaer A, Cordier Y, Roch O, Bischoff R. *Nucleic Acids Res.* 22: 3895, 1994
17. Apffel A, Chakel JA, Fischer S, Lichtenwalter K, Hancock WS. *Anal Chem.* 69: 1320, 1997
18. Easton DF, Ford D, Bishop DT. *Am.J Hum.Genet.*56: 265, 1995
19. De Stefano V, Chiusolo P, Paciaroni K, Casorelli I, Rossi E, Molinari M, Servidei S, Tonali PA, Leone G. *Blood.* 91: 3562, 1998
20. Jackson PE, Scholl PF, Groopman JD. *Mol Med Today.* 6: 271, 2000
21. Griffin TJ, Smith LM. *Trends Biotechnol.* 18: 77, 2000
22. Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, Hampel H, Markowitz A, Klimstra D, Jhanwar S, Winawer S, Offit K, Luce MC, Kinzler KW, Vogelstein B. *Nat Genet.* 17: 79, 1997
23. Kinzler KW, Vogelstein B. *Cell. 87*: 159, 1996
24. Su LK, Kinzler KW, Vogelstein B, Preisinger AC, Moser AR, Luongo C, Gould KA, Dove WF. [published erratum appears in *Science* May 22, 256(5060): 1114, 1992] *Science. 256*: 668, 1992
25. Torrance CJ, Jackson PE, Montgomery E, Kinzler KW, Vogelstein B, Wissner A, Nunes M, Frost P, Discafani CM. *Nat Med. 6*: 1024, 2000
26. Jackson PE, Groopman, JD. *Baillieres.Best.Pract.Res Clin Gastroenterol.* 13: 545, 1999
27. Qian GS, Ross RK, Yu MC, Yuan JM, Gao YT, Henderson BE, Wogan GN, Groopman JD. *Cancer Epidemiol.Biomarkers.Prev. 3*: 3, 1994
28. Hsu IC, Metcalf RA, Sun T, Welsh JA, Wang NJ, Harris CC. *Nature.* 350: 427, 1991
29. Bressac B, Kew M, Wands J, Ozturk M. *Nature.* 350: 429, 1991

30. Challen C, Lunec J, Warren W, Collier J, Bassendine MF. *Hepatology.* 16: 1362, 1992
31. Foster PL, Eisenstadt E, Miller JH. *Proc Natl Acad Sci.U S* A, 80: 2695, 1983
32. Puisieux A, Lim S, Groopman J, Ozturk M. *Cancer Res,* 51: 6185, 1991
33. Aguilar F, Hussain SP, Cerutti P. *Proc Natl Acad Sci.* U S A, 90: 8586, 1993
34. Cerutti P, Hussain P, Pourzand C, Aguilar F. *Cancer Res. 54*: 1934, 1994
35. Jackson PE, Qian GS, Friesen MD, Zhu Y-R, Lu P, Wang J-B, Kensler TW, Vogelstein B, Groopman JD. *Cancer Res. in press.* 2000
36. Rashid A, Wang JS, Qian GS, Lu BX, Hamilton SR, Groopman JD. *Br.J Cancer.* 80: 59, 1999
37. Kirk GD, Camus-Randon AM, Mendy M, Goedert JJ, Merle P, Trepo C, Brechot C, Hainaut P, Montesano R. *J Natl Cancer Inst.* 92: 148, 2000
38. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, Lockhart DJ, Morris MS, Fodor SP. *Science.* 274: 610, 1996
39. Little DP, Cornish TJ, O'Donnell MJ, Braun A, Cotter RJ, Koster H. *Anal Chem.* 69: 4540, 1997

CHAPTER 7

PROTEOMICS AND MASS SPECTROMETRY

Some aspects and recent developments

WV. Bienvenut, M. Müller, PM. Palagi, E. Gasteiger, M. Heller, E. Jung, M. Giron, R. Gras, S. Gay, PA. Binz, G J. Hughes, JC. Sanchez, RD. Appel, DF. Hochstrasser

University of Geneva Hospital, LCCC, Section R&D, Rue Micheli-du-Crest 24, CH:1211 Geneva 14, Switzerland. Tel:022-372-73-53; Fax:022-372-73-90; E-mail Denis.Hochstrasser@dim.hcuge.ch

1. INTRODUCTION TO PROTEOMICS

For several decades, DNA sequencing has progressed dramatically. Genomes from several bacteria, yeast and drosophila have been completely sequenced. Furthermore, the sequencing of the human genome is completed. In parallel, numerous genomic tools have been developed in order to study biological processes and explain physio-pathological findings in molecular terms. Indeed, the biological function of each gene should be understood. Therefore, after the stages of genome sequencing and gene discovery, attention must be focused on gene expression and the functions of the proteins they encode. DNA chip technology allows the simultaneous analysis of the expression of thousands of genes at the mRNA level and can unravel some biological processes. However, as previously demonstrated [1, 2], the correlation between the expression of mRNA and protein is low. In addition, many protein functions are related to their post-translational modifications such as phosphorylation or glycosylation and not to their level of expression. Consequently, large-scale studies of proteins or proteomes will be needed to complement genomic studies to better understand life processes. The word proteome was proposed by Marc Wilkins [3] to depict the PROTEin complement of a genOME. There are numerous proteomes for a single genome and proteomes are much more complex than genomes. Proteomics is the science which deals with the high throughput analysis of proteins, and this includes their identification, the measure of their level of expression and their partial characterisation. Thus, proteomics should complement genomics. Proteomics relies on efficient protein separation techniques, mass spectrometry, bioinformatics as well as gene and protein databases. One of the most powerful protein separation techniques is two-dimensional polyacrylamide gel electrophoresis (2-D PAGE or 2-DE gel) independently developed by Klose [4] and O'Farrel [5]. It has been further refined by the Andersons who proposed in 1975 the concept of a human protein index [6, 7]. It was certainly one of the early milestones

J.N. Housby (ed.), Mass Spectrometry and Genomic Analysis, 93—145.

of proteomics. Today, despite several drawbacks, 2-D PAGE is still very useful to analyse and display simultaneously thousands of proteins separated by charge and apparent size. In this chapter, new developments combining 2-D PAGE and mass spectrometry will be described. It will include the parallel chemical processing of proteins and the extensive use of bioinformatics tools and protein databases.

2. PROTEIN BIOCHEMICAL AND CHEMICAL PROCESSING FOLLOWED BY MASS SPECTROMETRIC ANALYSIS

Several years ago, identification of a single protein and its subsequent characterisation was a challenge for (bio-)chemists [8]. Historically, protein identification and characterisation was mainly conducted using the Edman degradation [9] in order to determine the primary sequence of a protein. Later on, N-terminal sequencing using Edman degradation was the method of choice to determine N-terminal or internal sequences which could be used to define an oligonucleotide probe specific for the mRNA encoding the protein in question. As Edman degradation provides a very accurate tool to determine a primary sequence, it is still in use. However, there are two major drawbacks. First, only 40 to 50 amino acids can be identified per day under normal conditions (Hughes G, unpublished results) and secondly, the N-terminal amino acid must be free of certain post-translational modifications, i.e. pyroglutamination or acylation to be available for Edman degradation. These days, this technique finds its application mostly in the characterisation of small proteins or peptide [10], in the quality control of recombinant proteins, in the determination of phosphorylation sites [11] or in the deduction of amino acid pairs that cannot be resolved by mass spectrometry, i.e., leucine/isoleucine, lysine/glutamine, phenylalanine/methionine sulfoxide, because of identical or nearly identical masses [12].

To overcome the low throughput characteristic of Edman degradation, amino acid composition analysis was implemented in the protein identification scheme. This method is based on the chromatographic analysis of free amino acids obtained after acid hydrolysis [13, 14, 15]. Amino acid analysis can achieve high throughput protein identification, however, there is a decrease in the confidence of identification [16, 17, 18]. A combination of amino acid analysis and Edman degradation, limited to 3-5 cycles to obtain a short sequence tag, was used to increase the confidence in protein identification [19, 20, 21].

At present, with the human genome nearly sequenced, whole proteome analysis presents new challenges for the identification and characterisation of the actual gene products, i.e. the proteins. While the genome represents a more or less unique set of data, the proteome is far more diverse as not all proteins are expressed at the same time and in the same tissues. Needless to say that such a huge project, probably involving the identification and characterisation of close to 1 million gene products is simply not feasible by Edman sequencing. This task needs accurate, reliable and rapid identification methods.

Two-dimensional electrophoresis gels (2-DE), a biochemical technique used to separate proteins according to their molecular weight and isoelectric point, emerged

in the middle 70's as a revolutionary procedure in protein analysis [4, 5]. Many of its technical problems had to be refined and it was only in the last 10 years that 2-DE gels have proven their capacity[22]. This revival was mainly due the combination of improved 2-DE techniques to mass spectrometer instrumentation, better computing and software tools, and the emergence of large protein sequence databases from genome-sequencing projects.

Peptide mass fingerprinting (PMF) involving analyses of peptides obtained after specific proteolytic digestion of polypeptides has shown its efficacy for protein identification [23, 24, 25, 26, 27). With the recent development of mass spectrometers (MS) such as Matrix Assisted Laser Desorption/Ionisation MS (MALDI-MS) [28] and the Electro-Spray Ionisation MS (ESI-MS) [29, 30], biopolymers can be efficiently measured and rapidly identified by database searches.

2.1. 2-DE gel protein separation

A critical problem in the post-genome era is the capacity of current techniques to perform large-scale separation of complex protein mixtures. A number of technologies have been investigated or are under development, such as capillary and gel electrophoresis, micro-channel networks [31] and liquid chromatography (LC). LC separation is starting to be widely used in proteomics. Link *et al.* have shown that multi-dimensional liquid chromatography coupled to a tandem mass spectrometer can analyse protein complexes [32]. They applied their approach to the yeast 80 S ribosome and identified ~100 proteins in a single run. Oda *et al.* [33] have described a MS-based method for simultaneous identification and quantitation of 2-DE separated proteins. Changes in post-translational modifications at specific sites on proteins were also determined. Accurate quantitation has been achieved by the use of whole-cell stable isotope labelling. Gygi *et al.* [34] have also published a method to quantify protein expression by using a new class of chemical reagents called isotope-coded affinity tags (ICATs) and tandem mass spectrometry (MS/MS). This ICAT technology provides a means to quantitatively compare global protein expression in semi-complex protein mixtures. However, it is still generally agreed, that two-dimensional gel electrophoresis remains unrivalled for its capacity to resolve several thousand polypeptides and to detect differentially expressed proteins. In 1970, Kenrick and Margolis published the first two-dimensional protein separation technique using native isoelectric focusing (IEF) and gradient gel electrophoresis [35]. However, the main tool used today to display and evaluate proteome complexity of any organism is the denaturing 2-DE independently developed by O'Farrell, Klose and Scheele in 1975 [4, 5, 36]. Many thousand polypeptides can be separated on the basis of different molecular properties in each of the two dimensions: charge (pI) in the first dimension and molecular mass (M_r) in the second. IEF is the electrophoretic technique in which the proteins are fractionated according to their isoelectric point (pI) through an immobilised pH gradient (IPG). This is achieved by using a set of weak acids and bases named Immobilines™. When the electric field is applied, only the sample molecules and any non-grafted ions migrate. Upon termination of electrophoresis, the proteins are separated into stationary isoelectric zones [22]. The majority of the laboratories

running 2-DE are currently using 3.5-10 or 4-7 IPG strips to display a wide range of proteins [37]. However, the use of 1 pH unit narrow range IPG permits higher protein loading and the investigation of a smaller, but more detailed "window" of the proteome. The work of Tonella *et al.* [38] demonstrated that the combination of 1 pH unit range gels with their high loading capacity (4 mg of proteins loaded) can display 85% of the *E. coli* proteome between the pI range from 5.09 to 6.09. After the first dimensional separation, the IPG strips are transferred onto classical vertical or horizontal SDS slab gels first published by Lämmli [39] to dissociate all proteins into their individual polypeptide chains. Thus, in addition to the analysis of the polypeptide composition of a sample, the investigator can also determine their apparent molecular mass

Today, silver staining[40] is probably the most popular non-radioactive protein detection as it is more sensitive than Coomassie Brilliant Blue [41] or reverse staining [42]. However, other procedures including fluorescent staining (Patton *et al.*, 43) and S^{35} or P^{32} radiolabelled samples can also be used [44]. Two-dimensional polyacrylamide gel patterns can be digitised and analysed on a computer to allow quantitative image analysis and automatic gel comparison by querying specialised protein databases. Several image analysis programs have been developed since 1975. The Swiss Institute of Bioinformatics (SIB) has developed and commercialised Melanie 3 (GeneBio S.A., Geneva), the design of which has been based on user friendliness. It allows the study of similarities and differences between sets of 2-DE images obtained from biological samples under different conditions (i.e. healthy vs. diseased or drug treated vs. non-treated samples) and could further help to find diagnostic, prognostic and therapeutic molecular markers in specific diseases or state [45].

2.2. Protein identification using peptide mass fingerprinting and robots

Thanks to the improvements in mass spectrometric technologies, MALDI-MS analysis has become a powerful tool for protein identification by the easy and quick PMF technique [23, 24, 25, 26, 27]. However, spot excision, protein digestion, peptide extraction and MALDI-MS sample preparation is a major bottleneck in the high throughput protein identification and characterisation process. A few groups proposed a robotic and/or computational approach to increase sample throughput, i.e. excision robots to cut out protein spots from 2-DE gels [46, 47], liquid handling systems to automatically pipette solvents [48, 49, 50] or computer programs to automate peak detection in mass spectra [51, 52].

In the standard manual procedure, protein spots of interest are chosen from the gel image. They are first excised from the gel and deposited in tubes or wells of a microtiter plate. After, they are individually subjected to an endoproteolytic cleavage step. At the end of this digestion, the sample is used for MS analysis. This manual procedure requires from the operator numerous and repetitive manipulations, needless to say that these repetitive steps are error prone.

Table 1. List of available high throughput protein identification systems

Company	*Internet URL address*	*Robot for sample preparation*	*MS machinery*	*Data handling system and related software*
Amersham Pharmacia Biotech	http://www.apbiotech.com	Under development	Ettan design LC-MS system (API-MS); ETTAN™ design MALDI-TOF	ImageMaster 2D Elite and Database. Partnership with Cimarron Software Inc.
Brucker Daltonics	http://www.bruker.com	MAPII	esquire3000, BiflexIII or ProflexIII MS	AutoXecute and MS Biotools supporting MASCOT
Micromass	http://www.micromass.co.uk	PROTEAN 2-D Spot Cutter, MassPREP™ Station	TOF Spec-2^E MALDI-TOF MS or Q-TOF MS,	Proteome Works System including PDQuest or Melanie II and MassLynx – ProteinLynx
Bio-Rad	http://www.bio-rad.com			
Genomic Solutions	http://www.genomicsolutions.com	Flexys Proteomics Robot, ProGest Protein Digestion Station and Pro-MS MALDI Prep Station		Investigator HT Analyzer Software
PE Corp	http://www.pecorporation.com http://www.appliedbiosystems.com	Symbiot workstation	Voyager STR MALDI-TOF MS, the Voyager DE-PRO or the Mariner API-TOF LC-MS	Protein Solution 1, SQL-LIMS, ProteinKeeper (under development)and Protein Prospector
Protana	http://www.protana.com	Home made: Robot for excision of gel spots; In-gel digestion system	Q*STAR, supplied by PE Sciex; REFLEX III, supplied by Bruker Daltonics	PPSS 2.2 (contains PepSea, ProteomeDB, Inspector, SoftSpot, PepSea FlowAgent)
ThermoQuest	http://www.thermoquest.com		Surveyor™ LC System, Finnigan LCQ™DUO, LCQ DECA and TSQ®	Xcalibur supporting TurboSEQUEST

A linear workflow includes the following steps:

1. Scan the gels;
2. Match the image with a master image or another image that will be used for comparison;
3. Choose the spots to be excised;
4. Cut out the spots, transfer them to vials (either tubes or microtitre plate wells);
5. Perform the required steps to digest the proteins in order to obtain peptide fragments that can be measured by MS;
6. Perform MS measurements;
7. Extract the MS raw data, treat and submit it to a protein identification software for database comparison.

During the whole procedure, a large amount of information is generated for each sample. This includes the description of both the gel and the chosen spot, the treatment of each spot, the vials in which they are processed, the MS files they generate and the identification results. There is a definite need for an automated process and this can be achieved using robots and laboratory management systems. Automation also decreases human work and as a consequence reduces possible human errors. A number of groups around the world are developing, optimising and integrating robotic systems to obtain what can be called a "robotised integrated proteomic solution". The ultimate goal is to set up a pipeline that includes the hardware and software components, e.g. a gel imaging system, a spot picker, a liquid handling system, a MALDI-plate loading system, a MALDI-MS instrument and a PMF identification tool.

An example of what is currently available is the ARRM-BR214 spot cutter distributed by Bio-Rad [47]. This robot takes an image of the gel or the PVDF blotted membrane, excises the spots and deposes them in a 96-well microtitre plate. There are more integrated systems available to perform multiple tasks. An example is the MultiprobeII robot from Packard Instruments which can perform a complete proteolytic digestion and load the peptide extract onto MALDI-MS sample plates. In addition to the control of their main tasks, these individual robots include other dedicated software components such as an image analysis program that automatically selects and excises spots of interest. However, there is no fully automated system available on the market. To this end, a number of collaborations and partnerships are actively working on these developments.

Depending on the required throughput, the overall data size to handle, the needed flexibility, laboratories might decide to choose one system or another, or to purchase only parts of a package. Table 1 is a non exhaustive list of systems that are currently under various phases of developments and commercialisation.

2.2.1. MALDI-MS analysis

Since the introduction of a new type of mass spectrometer [28, 53] capable of analysing intact proteins and polymers up to 100.000 Da, great interest has been shown for techniques allowing the analysis of proteins. Basically, the MALDI-MS technique consists of mixing the analyte (organic polymers, proteins or peptides)

with an excess of an organic compound (matrix) able to absorb the energy of a UV laser shot and to transfer this energy to the analyte for desorption and ionisation. Today, the preferred matrices are:

- α-cyano-4-hydroxy cinnamic acid (ACCA) [54, 55] for low molecular weight peptides (800-4500 Da);
- Sinapinic acid for larger polypeptides [56];
- Dihydroxy benzoic acid (DHBA) for glycopeptides and oligosaccharide [57, 58, 59].

Karas and Hillenkamp [28] used also nicotinic acid as a matrix and more recently Gusev *et al.* [60] proposed to improve the signal reproducibility by using fucose as a co-matrix with furulic acid and DHBA. To prevent the metastable fragmentation of glycoproteins, 3-hydroxy picolinic acid was proposed by Karas's group [61]. Other uncommon matrices were proposed such as 3-amino picolinic acid [62, 59], various trihydroxyacetophenone compounds [63], or meso-tetrakis(pentafluorophenyl)porphyrin [64]. More innovative propositions were made recently by Wei's group [65]. Here, the MALDI-MS sample-plate is the sample support and also the matrix. This surface is made of porous silicon chemically modified to achieve desorption and ionisation of the analyte directly from the surface without the intervention of an organic matrix.

During the first years of the MALDI-MS era, these spectrometers were mostly used to verify protein masses or to identify and characterise post-translational modifications [56]. In 1989, Henzel WJ, Stults JT and Watanabe W of Genentech Inc. presented the idea of protein identification using PMF in a poster at the Third Symposium of the Protein Society in Seattle. A second poster proposing this identification technique was shown in 1991 by Yates JR, Griffin PR, Hunkapillar T, Speicher S and Hood LE of Caltech during the Baltimore Symposium. Despite these 2 attempts, no real developments were done until 1993 when 5 groups [23, 24, 25, 26, 27] used this technique successfully to identify proteins. The comparison of *in silico* digested protein fragments with experimental masses became possible because of the development of computer programs [66, 67, 68] and computer facilities. At that time, only 4-5 peptides with an accuracy of 1-2 Da were generally enough to identify a protein compared to 5-10 peptides with an accuracy better than 50 ppm at present.

Success of this peptide analysis technique was due to:

- The production of singly charged ions;
- High sensitivity (far less material needed than for Edman degradation);
- A large mass range (from 500-600 Da up to a few hundred thousand Da);
- Short analysis time;
- Low sensibility to salts and contaminants,

Major drawbacks were the low resolution, a few hundred full-width half-maximum (FWHM) for ions above 10kDa, and the low accuracy of mass measurement. Ingendoh *et al.* [69] listed a series of causative factors such as broad initial energy distribution and spread in apparent generation time. They proposed modifications to improve peak resolution such as reducing the ion energy

distribution with an "ion reflectron" [70]. The ion reflectron is able to compensate for the flight time error due to energy spread which can reach up to 15 %. Also, a better peak resolution was obtained after focalisation of the laser beam (< 10 μm) and by the use of a 10^9 Hz digital oscilloscope. At that time the peak resolution was a few thousand FWHM for ions above 3000 Da.

Others groups [71, 72, 73, 74] obtained an improvement of mass accuracy and peak resolution by the use of a delayed extraction system [75]. This system uses a pulsed ion extraction MALDI ionisation technique increasing the accelerating voltage from 0 up to 3 kV in 300 nanoseconds. This technique allowed an increase of peak resolution for cytochrome c (12 kDa) from 350 FWHM obtained in linear mode to 1024 with a continuous ion extraction.

The major recent improvement in terms of mass accuracy and peak resolution is certainly the combination of the reflecting analyser with the delayed extraction mode [76]. They were able to achieve mass measurement accuracy of ±2 ppm on a 1000 Da peak with a resolution as high as 10.000 FWHM. Such improvement in accuracy has obvious implications in the reliability of protein identification using PMF. Higher resolution can be obtain when MALDI-MS is coupled with a Fourrier Transform Ion Cyclotron Resonance (FTICR) but the major drawback of such system is its expense.

Nevertheless, if existing instruments are able to provide a highly accurate mass determination, this technique is extremely dependent on matrices [61] and sample preparation [77]. Chen *et al.* investigated the interaction between the surface on which the sample is loaded and the peptide. They suggested that the surface, the solvent and the technique used to prepare the sample for MALDI-MS analysis influences ion signal.

Cohen *et al.* [78] and Figueroa *et al.* [79] investigated the influence of the solvent composition and the rate of matrix crystallisation. Cohen's group concluded that high molecular mass peptides must be prepared with a solvent that includes formic acid (solvent pH<1.8) and a slow crystallisation. On the other hand, for low M_r peptides, better results are obtained with trifluoroacetic acid/acetonitrile solvent (solvent $pH \approx 2$) and fast solvent evaporation. Figueroa's group also showed the importance of water concentration in the solvent.

Homogeneous co-crystallisation of the matrix and peptides/proteins is the critical step in sample preparation. Physical characteristics of a protein digest are widely distributed in terms of masses, pI and hydrophobicity. Kratzer *et al.* [80] found a preference of hydrophobic peptide adsorption to the non-polar (103) face. This fact was demonstrated by the crystallographic investigation of Beavis [81]. The speed of crystallisation can also explain the inclusion or exclusion of hydrophilic peptide into the growing crystal. The secondary structure of a peptide will also affect the signal intensity. Wenschuh *et al.* have shown [82] that MALDI-MS signal response of peptides displaying stable α-helical and β-sheet structures was different when two adjacent amino acids were replaced by their corresponding D-isomers. A simple D-L amino acid modification may disrupt α-helical and β-sheet structures and therefore completely alter the MALDI-MS spectral pattern.

Slow crystallisation of the matrix helps for co-crystallisation with predominant adsorption of hydrophobic peptides in large crystals. In this case, hydrophobic peptides occupy the adsorption sites on the (103) crystallographic face and this produces a stable architecture. If solvent is quickly evaporated (flash evaporation), matrix micro-crystals are obtained. Peptide adsorption on the hydrophobic adsorption sites is controlled by the kinetics of analyte diffusion to the adsorption sites on the (103) crystallographic face. This type of co-crystallisation is thermodynamically less stable than the previous one due to the higher potential energy. In this case, hydrophobic and hydrophilic peptides can be integrated in the crystal structure without site competition allowing a better distribution and decreased discrimination between them [83]. This behaviour can also explain the suppression effect reported in the literature [84, 78]. In a few articles, the use of surfactants was proposed to decrease the suppression effect. N-Octylglucoside [78] showed a positive effect on high M_r peptides but anionic surfactants [85] were preferred to analyse hydrophobic peptides. The use of surfactants helps to create a more homogeneous distribution of hydrophobic and hydrophilic peptides integrated in the matrix by decreasing the stabilising hydrophobic effects between peptide and matrix.

Krause [86] found that 94% of the most intense peaks bore an arginine (R) residue at the C-terminal side of the tryptic fragment obtained from digested proteins from mycobacteria. They obtained higher signal intensities for peptides containing R than for those with lysine (K) at the C-terminal amino acid. They attributed this effect to specific chemical properties of R by the comparison of signal intensity with similar peptide carrying K or R at the C-terminal. Meanwhile, an exhaustive study conducted by Keil [87] on the endoproteinase specificity concludes to a lower activity of trypsin towards lysyl compared to arginyl residues. In this case, after tryptic digestion, C-terminal R peptides would be at higher concentration in the protein digest, thus explaining the higher intensity of these peaks.

The above facts do not favour protein or peptide quantitation using MALDI-MS. Some problems are associated with MALDI-MS quantification: i) low shot-to-shot reproducibility, ii) various signal suppression effects, and iii) strong influence of sample preparation and matrix crystallisation. Nevertheless, it is possible to use MALDI-MS to obtain absolute or relative quantitation. In most cases, the idea is to use an internal standard for an absolute quantitation, but this standard must have the same physico-chemical characteristics as the quantified peptide. The use of a different peptide in terms of sequence may result in different desorption and ionisation properties. Usually, the internal standard is the same peptide labelled with a stable isotope to modify slightly the mass.

Gobom *et al.* [88] developed a method to quantify neurotensin in human brain tissue. For a 10 shots cumulative spectrum, they obtained more than 20% variation of the signal intensity. Due to the low reproducibility of shot-to-shot signal intensity, this technique needed up to 400 cumulative acquisition to minimise this problem. Under those conditions, they obtained a variation of ±2%. In this case, MALDI-MS as a quantitation technique was not as good as the reference method but allowed more specific information to be obtained.

Hensel *et al.* [89] proposed to use an electrospay method to prepare the sample. With this technique, they were able to decrease the coefficient of variation more than 3 times as compare to air-dried samples. Gygi *et al.* [34] proposed a relative quantitation of all proteins contained in a sample using a special alkylating agent called isotope-coded affinity tag (ICAT). This technique was developed for MS/MS analysis but it can also be used with MALDI-MS.

To conclude, MALDI-MS has been greatly improved since Karas' and Tanaka's first descriptions[28, 53]. It is now possible to obtain routinely a peak resolution better than 2000 FWHM and mass accuracy below 30 ppm, and under these conditions, most of the digested proteins can be clearly identified. Unfortunately, some proteins cannot be directly identified by this method and more information about their primary structure is required. Such information can be obtained by MS/MS techniques or by specific chemical modification as described below.

2.2.2. MS/MS analysis

PMF can sometimes give ambiguous results: if the PMF results have to be searched against large sequence databases; if the peptides have post-translational modifications; if the sequence of the protein under investigation is not known. Then, obviously, more sequence information is needed. Spectra acquired from fragmented peptides either by post-source decay (PSD) or collision induced dissociation (CID), can be used to determine sequence tags or the complete sequence of a peptide, with the help of computer algorithms. This additional peptide sequence information makes protein identification less ambiguous and can be used to search expressed sequence tags (EST) databases in case the protein is not yet listed in a protein database.

MALDI-RETOF-PSD MS analysis

PSD of peptides relies on the metastable decay of ions in the first field-free drift tube of a time of flight (TOF) analyser. Metastable decay is initiated by low-energy collisions of neutral matrix molecules (dominant in the desorption plume) with ionised analyte molecules during the initial stage of acceleration. The different masses of the fragment ions are separated by dispersion of the in-time ions with different kinetic energies in the electrostatic reflector field. These are detected after the second field free drift tube [90, 91]. As mentioned above, MALDI-PSD relies heavily on metastable ions and the most common PSD fragment ions produced are a,b,y,z and d ions according to the defined nomenclature [92, 93]. These ions could also produce satellite ions that loose ammonia or water [94]. Direct utilisation of MALDI-PSD MS for unknown peptides sequencing is not easy and generally not very sensitive. In contrast to CID with a collision gas in a collision cell, there is little control on the degree of dissociation of the reaction pathways and the fragments produced.

Chemical modification of the peptides prior to PSD MS measurement was proposed to improve sequence identification [95], to facilitate peptide specific fragmentation and to suppress a part of the spectrum in order to simplify it. Pappin,

Spengler and Allison's groups [96, 97] used a modified N-terminal amino acid group with a quaternary ammonium ion. This charged group at one end of the peptide facilitates fragmentation and allows much simpler spectra with mostly a type ions.

Lacey's group [98, 99, 100] modified the N-terminal end of the peptide with a negatively charged compound. In this case, y-ion type fragmentation is seen, a, b and c types of ions are suppressed due to the negative charge carried by the sulfonate on the N-terminus.

Vandekerckhove's group developed an interesting technique to obtain MALDI-PSD spectra [101-103]. POROS R2 beads were used to extract and concentrate diluted peptides from the protein digest. Then, a washing step was done to remove contaminating salts. The whole sample, POROS R2 beads and peptides, were mixed with the matrix and loaded on the MALDI sample plate for PSD analysis. With this method, mostly y-ions were produced. Nevertheless, in a review dealing with protein identification methods, Gevaert and Vandekerckhove recognised [104] that sequence determination for an unknown sequence was difficult using this technique and that *de novo* sequencing was impossible.

To overcome the limitations inherent with PSD peptide sequencing using MALDI instrumentation, considerable research efforts are under way to use a MALDI source in conjunction with tandem MS and a CID cell. One approach is the MALDI-TOF/TOF as a linear configuration [105, 106, 107]. Another solution is the use of the more mature orthogonal acceleration TOF technique as the second stage mass analyser with either a quadrupole parent ion selector [108, 109] or a linear TOF separator upfront to the CID cell [110]. These are rather new developments in the field of tandem MS. The most common and furthest developed ionisation method in conjunction with CID, is electrospray ionisation (ESI).

ESI-MS/MS analysis

Mass spectral analysis requires that the analyte is introduced into the mass spectrometer as an gaseous ion. This is a major hurdle especially for the ionisation of biological molecules, which consist mostly of large and therefore extremely non-volatile polymeric units. Nevertheless, several ionisation methods were developed during the last decades. Among them, MALDI and ESI were the most successful because of their high ionisation efficiencies, i.e. very high ratios of (molecular ions produced)/(molecules consumed) [90]. A major advantage of ESI is that it produces multiply charged ions in an almost linear correlation of charges added per mass increment of a polypeptide [111]. Thus, molecular ions can be analysed based on their mass-to-charge ratio (m/z) which is mainly in the range of m/z = 500-3000. As a result, the mass analyser can be kept relatively simple and mass measurements are very precise. These advantages were recognised independently by two research groups in the 1980's, Yamashita and Fenn [29] in the US and Aleksandrov *et al.* [30] in the former USSR. The latter group also accomplished for the first time the on-line coupling of liquid chromatography to a mass spectrometer [112] which is still one of the strengths of ESI.

The actual mechanism of the ESI process is still a matter of debate and the dedicated reader may refer to the literature, e.g. Kebarle and Peschke [113]. In

summary, the analyte solution is pumped through a narrow bore capillary held at a potential of a few kilovolts relative to a counter-electrode situated normally behind a first set of ion focusing lenses. This results in a fine mist (spray) of small charged droplets under atmospheric pressure. The charge on the droplets drives them through the inlet orifice and sampling lenses/capillaries under differential pumping into the high vacuum system for mass analysis. The use of a warmed drying gas, gradually reduced pressure and heating, desorbs the charged droplets. The droplets go through a cascade of so-called Coulomb explosions initiated because droplet shrinking results in critical Rayleigh diameters. This process leads finally to the production of completely desorbed and multiply charged analyte ions. The focused ion beam is then subjected to a first mass analyser, which is in most cases a quadrupole, using radio frequency signals to scan through the m/z range or select an ion with a specific m/z to be analysed. ESI is a soft ionisation method resulting generally in no fragmentation of the sample ions. However, ions of a specific m/z value separated in the first mass analyser can be fragmented in a collision cell in the presence of a few mTorr of a neutral and inert gas. This low energy collision-induced dissociation (CID) process is very efficient on doubly or triply charged peptides inducing mainly fragmentation of amide bonds in the peptide backbone. Commonly, tryptic digests are used for this type of analysis and on average peptides retain two positive charges when ionised in positive ion-mode, one on the basic N-terminal and the second one on the basic C-terminal amino acid K or R, respectively. Thus, the product ions are generally singly charged and contain either the intact N- or C-terminal amino acid of the b- or y-ion type, respectively. The masses of the fragment ions are then measured in the second mass analyser of a tandem mass spectrometer. The resulting MS/MS spectrum contains sufficient information to deduce sequence tags of the fragmented peptide or can be automatically correlated with sequence databases for rapid and reliable protein identification (see section 3.2 below). Fast instrument controlling software can switch quickly the tandem mass spectrometer from MS to MS/MS mode thus enabling automated data acquisition of the fragments produced.

Probably the simplest configuration of a tandem mass spectrometer is the so-called triple quadrupole set-up, consisting of a first quadrupole to scan parent ions, a second quadrupole which serves as the CID cell, and a third quadrupole used to scan product ions. This type of tandem mass spectrometer was the first to be commercially available. Although having a rather limited mass resolution of around 1000-2000 $(m/\Delta m)$ these instruments are still used in many labs around the world due to the possibility of doing single ion monitoring and parent ion scanning. A more recent implementation of tandem MS was already described in the previous section, where product ions produced in the CID cell are measured in an orthogonal acceleration TOF compartment [114]. These instruments combine ESI-quadrupole technology with the superior mass resolution and sensitivity of a TOF analyser. A resolution of 5000 is standard with a Q-TOF instrument. Another type of instrument with a similar resolution as the triple quadrupole MS relies on a different technique by collecting ions in a potential trap. The ions can be scanned in MS mode by altering the radio frequency amplitude of the trap, which leads to the ejection of the ions into the detector. For MS/MS, the trap is filled with ions of a specific m/z value

by adjusting the RF amplitude followed by introduction of the collision gas and scanning of the fragments. The great advantage of an ion-trap MS is its operation speed, which is up to ten times faster than a quadrupole instrument at identical sensitivity.

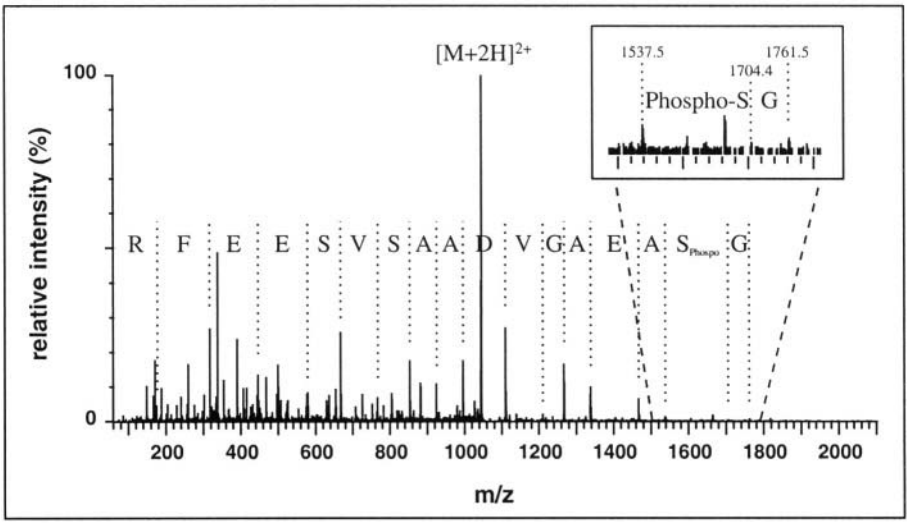


*Figure 1. A tryptic digest of chicken ovalbumin was passed over an immobilised metal affinity column (IMAC) to isolate phosphorylated peptides. The phosphopeptide enriched eluate in 0.1 M sodium phosphate buffer was desalted with ZipTip and analysed with nano-ESI-MS/MS on a Q-Tof (Micromass, Manchester, UK). The peak at m/z = 1044.95 recorded in the MS survey scan was induced to fragmentation by collision with Argon gas and the recorded MS/MS spectrum was subjected to interpretation with SEQUEST, searching against a July 2000 release of SWISS-PROT. The peptide was identified as EVVGS*AEAGVDAASVSEEFR from ovalbumin with the serine residue at position five modified by a phosphate ester group. Peaks corresponding to the y-ion series of the underlined part of the above sequence were found in the spectrum denoted by dotted lines and single letter amino acid symbols in the figure. The inset represents a zoomed-in region of the spectrum showing the sequence phosphoserine-glycine.*

As mentioned above, ESI instruments were coupled to liquid chromatography. The biggest impact in ESI MS has been the adaptation to reduced flow capabilities in the 10-500 nl/min range. Wilm and Mann [115] and Emmett and Caprioli [116] developed such improvements in parallel. The combination of nano-flow LC with micro-capillary reversed phase HPLC and nano-ES has i) dramatically improved the sensitivity of ESI-MS/MS and ii) enabled the automation of protein identification by using an auto sampler for loading of samples onto the LC [117].

In Figure 1, an example of a MS/MS spectrum is shown illustrating one of the strengths of ESI-MS/MS, namely the characterisation of post-translational modifications of proteins. The recorded spectrum of a doubly charged peptide with m/z = 1044.95 contained all the information to identify and characterise

unambiguously the phosphorylated peptide. Although the singly phosphorylated peptide with the sequence shown in Figure 1 contains three serine residues as potential phosphorylation sites, the fragment ion pattern of the y-ion series demonstrated that only serine in position five could be phosphorylated. Indeed, this serine residue is a known phosphorylation site of chicken ovalbumin. The interpretation of this spectrum was greatly facilitated by the use of SEQUEST (see section 3.2), which scored this particular spectrum with a relatively high cross-correlation score.

2.2.3. Improvement of the identification by chemical modification of peptides

The expansion of protein sequence databases, e.g. TrEMBL, SWISS-PROT, NCBInr brought about by genome sequencing projects, decreases the probability of obtaining an unequivocal protein identification by PMF alone [118]. More information like amino acid sequence or amino acid composition increases the confidence in any protein identification. Lahm's group [118] defined three ways to reduce this problem:

- Acquisition of an other MALDI-MS spectra with optimised parameters,
- Use of a different endoproteinase to generate a different PMF,
- Identification of a short sequence tag using MALDI-PSD and/or ESI-MS/MS.

The time required for such experiments prevents such methods being used for high throughput protein identification [119].

Specific modification of one or more amino acids in a given peptide chain could more easily supply the necessary information. Possible chemical reactions must have the following requirements:

- They must be fast;
- Have a high yield of conversion;
- Be simple to handle;
- Use reagents and buffers that do not leave by-products which may alter matrix crystallisation and ionisation process;
- Ideally, reaction should be done on the MALDI sample plate.

Alternatively, reactions should be done with a sample already embedded in matrix on the MALDI-MS sample plate. In the following section. Two techniques, involving the modification of samples are described below [120].

Esterification

A well-documented modification is the esterification of side chain carboxylic groups of. glutamic (D) and aspartic (E) acids together with the carboxy-terminal group. Usually, methanol is used as the methylation reagent [121, 122, 123, 84, 124, 125]). Other alcohols can be used, e.g. 2-propanol, 1-butanol, 1-hexanol, 1-octanol, benzyl alcohol [126] or ethanol [127]. This type of treatment allows the determination of the number of D and E in a peptide, thus increasing the confidence of its identification.

The technique used in our laboratory is an adaptation of Pappin's method [124]. Briefly, the esterification reagent is obtained by addition of thionyl chloride to dry methanol at –80°C to form a 1% solution. Fresh reagent (10µl) is added to an Eppendorf tube containing the dried peptide sample After incubation at 55°C for 20 minutes, the excess reagent is removed by vacuum centrifugation. The modified peptides are re-suspended in 5 µl of acetonitrile/water/TFA solution (50:49:1 v:v:v). Two µl of the sample are loaded on the MALDI sample plate prior to matrix addition and MALDI-MS analysis [120].

Table 2. Peptides identified in the tryptic digest of native ovalbumin. Bold characters in the sequence column represent potential esterification sites. Last column (Ester) shows if the peptide was also found in the esterified sample listed in Table 4 ('+' if found, '-' otherwise)

Peptide mass	*Position*	*Sequence*	*Ester*
1209.52	190-199	**DED**TQAMPF**R**	+
1345.738	370-381	HIATNAVLFF**GR**	-
1465.776	111-122	YPILP**E**YLQCV**K**	-
1555.721	187-199	AFK**DED**TQAMPF**R**	+
1571.716	187-199	AFK**DED**TQAM*PF**R**	+
1581.721	264-276	LT**E**WTSSNVM**EER**	+
1597.716	264-276	LT**E**WTSSNVM***EER**	+
1687.84	127-142	GGL**E**PINFQTAADQA**R**	+
1773.899	323-339	ISQAVHAAHA**E**IN**E**AG**R**	+
1858.966	143-158	**E**LINSWV**E**SQTNGII**R**	-
2008.946	340-359	EVVGSA**E**AGV**D**AASVS**EEFR**	+

To perform a comparison of the untreated and treated samples, both are loaded on the MALDI sample plate and analysed (Figure 2). The spectrum of the unmodified digest is used for a primary protein identification using PMF (Table 2). The modified material is also analysed by MALDI, and masses of the modified peptides are used for protein identification using PMF with the Mascot program (see section 3.1.2.). In this case, esterified carboxylic groups are considered as permanent modifications like chemical modifications of cysteines, e.g. carboxymethylation.

The comparison of the two peak lists allows a determination of potentially esterified peptides and their degree of esterification. The mass difference of "n x 14.0157" (the mass difference induced by the esterification) is checked between the peak list of the untreated and treated samples. Results are summarised in Table 3.

Protein identification based on the PMF mass lists of the treated and untreated samples mostly resulted in the correct identification but with a non significant score (Table 5). The combined use of the mass list of untreated sample and the corresponding degree of esterification listed in Table 3 allows a clear identification of the correct protein with a highly significant score.

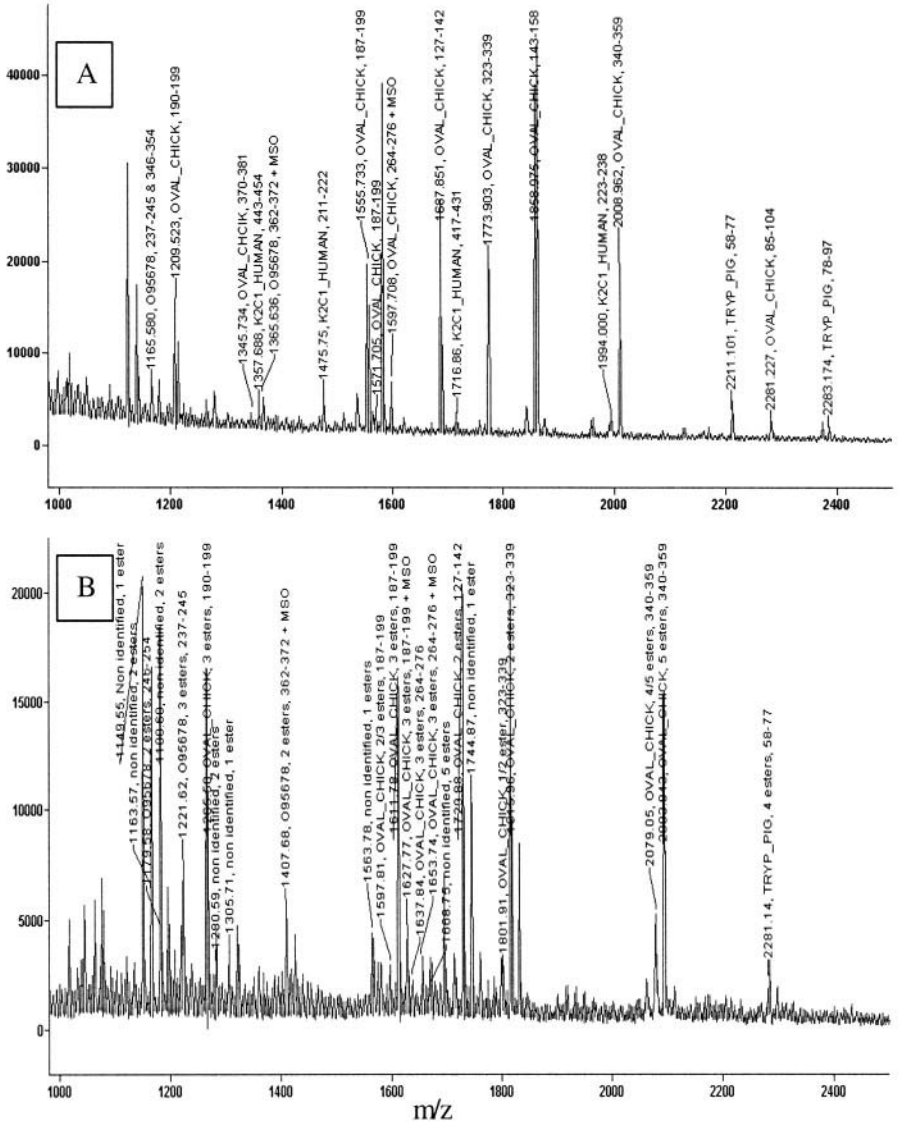


Figure 2. A) MALDI-MS spectrum of the native OVAL_CHICK digest peak labels show peptide mass, peptide sequence, protein source and the peptide position within the protein. B) MALDI-MS spectrum of the esterified sample. Peak labels show peptide mass, protein source, experimental number of esterifications, and the peptide position within the protein.

Table 3. Peak list comparison. List of masses showing a difference of "n x 14.0157 Da" with less than 15 PPM error between MALDI-MS spectra from untreated and esterified samples. Init. mass: list of masses in the native sample spectrum that matched with peak masses of the esterified sample. Mass ester: masses in the esterified sample spectrum; # ester: potential number of ester groups on the peptide chain (n: from 1 to 6); # D & E: potential number of D and E amino acids in the peptide sequence (n-1 due to the C-terminal esterification); Protein source: the protein to which the peptide corresponds (Protein named O95678 and K2C1_HUMAN are two cytoskeletal keratin type II from hair); Bold letters in sequence column represent potential position of the esterification site on the sequence from D and E amino acids and letters in italics represent potential esterification at the C-terminal amino acid. MSO: methionine sulfoxide

Init. Mass	*Mass ester*	*# ester*	*# D & E*	*Protein source*	*Sequence*
1121.53	1149.56	2	1	Unknown	
1121.53	1163.58	3	2	Unknown	
1138.55	1180.60	3	2	Unknown	
1165.56	1179.59	3	2	O95678	Y**EE**LQVTAG*R*
1165.58	1221.64	4	3	O95678	VRY**EDE**IN*K*
1179.58	1221.64	3	2	K2C1_HUMAN	**D**YQ**E**LMNT*K*
1209.52	1265.58	4	3	OVAL_CHICK	**DED**TQAMPF*R*
1277.67	1305.70	2	1	Unknown	
1365.64	1407.68	3	2	O95678	NTKQ**E**IS**E**MN *R*
1535.75	1563.78	2	1	Unknown	
1555.73	1597.80	3	2	OVAL_CHICK Partially esterified	AFK**DED**TQAM PF*R*
1555.72	1611.78	4	3	OVAL_CHICK	AFK**DED**TQAM PF*R*
1571.72	1627.78	4	3	OVAL_CHICK MSO	AFK**DED**TQAM PF*R*
1581.73	1637.80	4	3	OVAL_CHICK	LT**E**WTSSNVM **EE***R*
1584.68	1668.76	6	5	Unknown	
1597.71	1653.75	4	3	OVAL_CHICK MSO	LT**E**WTSSNVM **EE***R*
1687.84	1729.89	3	2	OVAL_CHICK	GGL**E**PINFQT AA**D**QA*R*
1716.86	1744.89	2	1	Unknown	
1773.87	1801.90	2	1	OVAL_CHICK Partially esterified	ISQAVHAAHA **EINE**AG*R*
1773.91	1815.96	3	2	OVAL_CHICK	ISQAVHAAHA **EINE**AG*R*
2008.97	2079.04	5	4	OVAL_CHICK Partially esterified	**E**VVGSA**E**AGV DAASVS**EEF***R*
2008.95	2093.04	6	5	OVAL_CHICK	**E**VVGSA**E**AGV DAASVS**EEF***R*
2211.06	2281.14	5	4	TRYP_PIG	LG**E**HNI**D**VL**E** GN**E**QFINAA*K*

Table 4. Peptides identified from the peak list of the esterified sample. Mass ester: masses in the esterified sample spectrum; Init. mass: masses in the native sample spectrum; # ester: number of ester groups on the peptide chain; Seq. Pos.: sequence position; Sequence: amino acid sequence of the corresponding peptide; Bold characters represent position of the esterification site in the sequence and M: methionine sulfoxide*

Mass ester	*Init. mass*	*# ester*	*Seq. Pos.*	*Sequence*
1729.889	1687.840	3	127-142	GGL**E**PINFQTAADQA**R**
1815.957	1773.899	3	323-339	ISQAVHAAHA**E**INEAG**R**
1265.578	1209.520	4	190-199	**DED**TQAMPF**R**
1611.783	1555.721	4	187-199	AFK**DED**TQAMPF**R**
1627.782	1571.716	4	187-199	AFK**DED**TQAM*PF**R**
1637.796	1581.721	4	264-276	LTEWTSSNVM**EER**
1653.747	1597.716	4	264-276	LTEWTSSNVM*EER
2093.043	2008.946	6	340-359	**E**VVGSA**E**AGV**D**AASVS**EEFR**

Table 5. Identification results for chicken ovalbumin (OVAL_CHICK) and score using Mascot PMF tool. Mass error is limited to 15 PPM, M_r of the protein is fixed to 45 kDa, methionines could be oxidised, cysteines are native and the NCBInr database was used. Research for protein identification was conducted against all entries contained in the database and a second time only against "lobe-finned fish and tetrapod clade" to reduce the size of the database. Id. Rank: OVAL_CHICK identification rank; Sc.: Score; Stat.: significance of the result at P<0.05, Scores are significant for values higher than 62 (see section3.1.2. the description of Mascot tool). Id. Pept.: number of identified peptides in the PMF output

	All NCBInr entries				*Lobe-finned fish and tetrapod clade*			
Sample	*Id. rank*	*Sc.*	*Stat.*	*Id. Pept.*	*Id. rank*	*Sc.*	*Stat.*	*Id. Pept.*
Native sample	1	60	-	11	1	60	-	11
Esterified sample	2	32	-	8	1	32	-	8
Combined results: Table 3	1	111	+	8	1	111	+	8

H/D exchange: Quantitation of labile protons on peptides

Hydrogen/deuterium (H/D) exchange is a common practice in biochemistry. Numerous articles have described the exchange of protein labile hydrogen's using heavy water (D_2O) and other deuterated solvents, e.g. D_4-methanol (MeOD), D_3-acetonitril (D_3-AcN), D_1-trifluoroacetic acid (D-TFA). This technique is mostly used to identify specific binding sites [128], conformational changes [129, 130, 131] and deduce secondary structure of proteins [132, 133]. Proteins are usually incubated in a deuterated solvent and labile protons are exchanged with different kinetics during the incubation step. Hydrogen/hetero-atom binding energy and labile hydrogen protein steric positions are the most important factors in the kinetics of the reaction.

For example, in a protein complex involving two agonists, binding and/or adsorbing sites are differently exposed to the solvent, and these hydrogen's therefore have a different kinetic rate of exchange [134]. For a β-sheet, hydrogen's involved in the kinetic of H/D exchange show a lower rate of modification that is correlated with the stabilisation of the hydrogen's due to the secondary structure [132]. These modifications or kinetics of exchange could be visualised using nuclear magnetic resonance or quantified using MS. This information is used to reconstruct the 3D-protein configuration.

Table 6. Number of exchangeable hydrogen's in common amino acids

Amino acid	*1 letter code amino acid*	*# of exchangeable protons*
Alanine	A	1
Arginine	R	5
Asparagine	N	3
Aspartic acid	D	2
Cysteine (native)	C	2
Cysteine (carbamidomethyl)	C-CAM	3
Cysteine (propionamide)	C-PAM	3
Glutamic acid	E	2
Glutamine	Q	3
Glycine	G	1
Histidine	H	2
Isoleucine	I	1
Leucine	L	1
Lysine	K	3
Methionine	M	1
Phenylalanine	F	1
Proline	P	0
Serine	S	2
Threonine	T	2
Tryptophane	W	2
Tyrosine	Y	2
Valine	V	1

Spengler *et al.* [95] proposed to use H/D to exchange all the labile hydrogen's of a peptide. The comparison by MS of the native and H/D exchanged peptides allowed a determination of the number of labile protons carried by the peptide. This technique was shown to have two advantages. First, they used this type of information during MALDI-PSD analysis to confirm the identified amino acids with the number of exchanged protons (see Table 6). Comparison of the PSD spectrum of treated and non-treated peptides facilitated the interpretation of the peptide amino acid sequence [95, 135]. Second, direct MALDI spectra comparison of the native and treated peptides permitted the determination of the number of exchangeable protons. When this technique was applied to the whole digest of a protein, such

information was used to confirm or reject the previously proposed sequence using PMF.

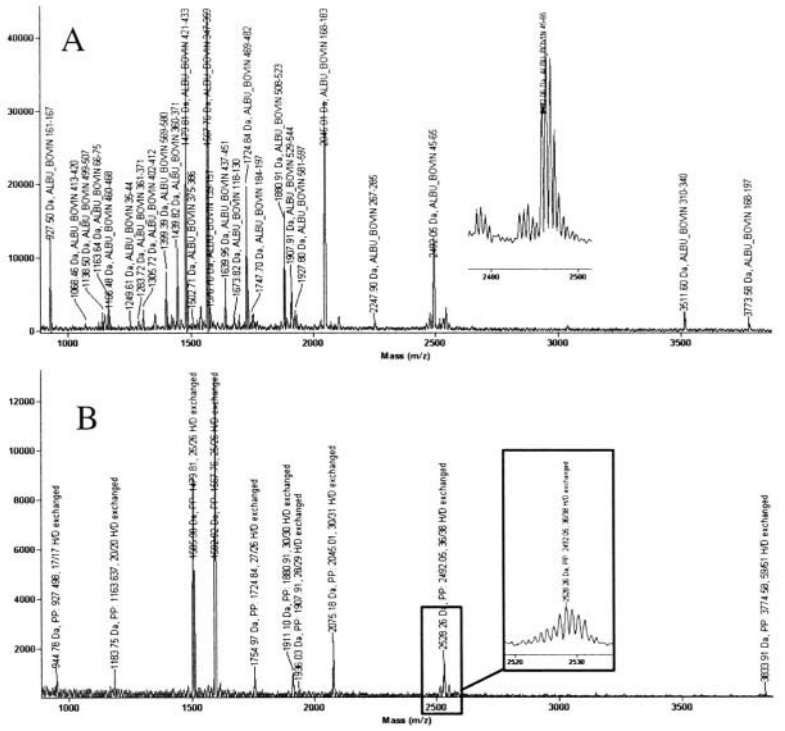


Figure 3. A) MALDI-MS spectrum of bovine albumin digest: Peak annotations correspond to the peptide M_r, the name of the protein and the peptide position in the sequence. B) MALDI-MS spectrum of the same sample after treatment with deuterated solvent: peak annotations correspond to the peptide M_r after treatment, Previous Peptide (PP) M_r and the number of H exchanged compared to the theoretical number.

To obtain reliable data, the H/D exchange must be as complete as possible. Also, the quality and the composition of the deuterated solvent used are crucial. Figueroa *et al.* [79] showed that the concentration of D_2O in these solvents is really important for the H/D exchange equilibrium. For example, a low percentage of D_2O (1-3 % in deuterated methanol) showed equilibrium at 77% H/D exchange in bradykinin. When D_2O concentration was increased to 40%, the exchange reached its maximum for an equilibrium exchange of 97.2%. Composition of the solvent is also important since the reaction is mainly limited by kinetics. For this type of reaction where labile protons are replaced by deuterium, free D^+ ions in the solvent could be considered as catalyst for the exchange reaction. Also, the remaining quantity of hydrogen's in the deuterated solvent considered as contaminants is competing in the exchange

reaction. To limit the competition effect of H with D in the exchange reaction, such solvents must be of the highest quality grade.

Practically, after acquiring a first MALDI-MS spectrum of the native sample, the matrix/analyte crystals are treated with a solution of MeOD/D_2O/TFA (70:30:1, v:v:v). A volume of two µl of the deuterated solvent are added to the sample in a closed-box flushed with dry nitrogen. After 30 seconds of incubation, the remaining solvent is evaporated under vacuum. This step is repeated 3-4 times before a new MALDI-MS spectrum is acquired. Labile hydrogen exchange is usually around 90 to 95%. Due to this partial exchange, treated peptide masses show a larger isotopic distribution after deuteration (Zoom boxes in Figure 3). Therefore, the highest peak of the distribution is used to define the average mass of the modified peptide for further calculations.

Figure 3 shows an example of protein treatment with deuterated solvent for H/D exchange. The bovine albumin digest (Figure 3A) was treated directly on the MALDI sample plate as described above and a second spectrum was acquired (Figure 3B). Results of the PMF identification on the first spectrum querying the SWISS-PROT and TrEMBL databases for mammalian species allowed the identification of bovine albumin (ALBU_BOVIN, P02769) at the first rank and Yellow mealworm ecdysone receptor (O02035). In the case of bovine albumin, 10 peaks out of 22 were matched in both spectra (Table 7).

The average exchange of hydrogen to deuterium on the 10 identified peptides corresponded to 97.7 ± 2.0 %. The second scoring protein in the output of the PMF tool was identified only with 3 peaks out of 9 (Table 8) and then the exchange rate was calculated to be 86.2 ± 8.0 %. Thus, this technique has a clear discriminating power enabling the unambiguous identification of a protein. Additionally, this technique has the major advantage that the same sample prepared for protein identification by PMF is reused to do H/D exchange. The technique is prone for automation and therefore large numbers of samples can be analysed in a high throughput mode.

2.3. *The molecular scanner approach*

Well known high throughput approaches combine two-dimensional electrophoresis (2-DE) with PMF analysis [136, 137]. Although automation is often possible, a number of limitations still adversely affect the rate of protein identification and annotation in 2-DE databases:

- The sequential excision process of pieces of gel containing protein; the enzymatic digestion step,
- The interpretation of mass spectra (reliability of identifications),
- The manual updating of 2-DE databases.

Methods involving high resolution protein separation, paralleled sample preparation, automation of experimental processes and of database comparison, as well as powerful and specific visualisation tools need to be developed and integrated [138, 3].

Tables 7 and 8: Identified peptide from ALBU_BOVIN (Table 7) or O02035(Table 8) before and after treatment with deuterated solvent. Init. Mass: identified peptides from ALBU_BOVIN from untreated spectrum; Seq. Pos.: position of the peptide in ALBU_BOVIN sequence; Sequence: sequence of the corresponding peptide; Theo. H/D: Theoretical number of exchangeable hydrogen's; H/D Mass: peptide mass after sample treatment; Exp. H/D: experimental number of exchanged hydrogen's; % H/D: % of hydrogen's exchange; NV: peak not visible after treatment

Init. mass	*Seq. Pos.*	*Sequence*	*Theo. H/D*	*H/D Mass*	*Exp. H/D*	*% H/D*
927.50	161-167	YLYEIAR	17	944.78	17	100
1068.46	413-420	QNC*DQFEK	22	NV		
1163.64	66-75	LVNELTEFAK	20	1183.75	20	100
1166.49	460-468	C*C*TKPESER	25	NV		
1249.61	35-44	FKDLGEEHFK	21	NV		
1283.72	361-371	HPEYAVSVLLR	21	NV		
1305.72	402-412	HLVDEPQNLIK	22	NV		
1399.70	569-580	TVMENFVAFVDK	22	NV		
1439.82	360-371	RHPEYAVSVLLR	26	NV		
1479.81	421-433	LGEYGFQNALIVR	26	1505.98	26	100
1567.76	347-359	DAFLGSFLYEYSR	26	1592.92	25	96.2
1576.76	139-151	LKPDPNTLC*DEFK	26	NV		
1639.95	437-451	KVPQVSTPTLVEVSR	29	NV		
1724.84	469-482	MPC*TEDYLSLILNR	27	1754.97	26	96.3
1747.70	184-197	YNGVFQECCQAEDK	31	NV		
1880.91	508-523	RPCFSALTPDETYVP K	30	1911.10	30	100
1907.91	529-544	LFTFHADIC*TLPDTE K	29	1936.03	28	96.6
1927.80	581-597	C*C*AADDKEAC*FAVEG PK	33	NV		
2045.01	168-183	RHPYFYAPELLYYAN K	31	2075.18	30	96.8
2247.90	267-285	EC*C*HGDLLEC*ADDRA DLAK	41	NV		
2492.05	45-65	GLVLIAFSQYLQQC*P FDEHVK	38	2528.26	36	94.7
3774.58	168-197	RHPYFYAPELLYYAN KYNGVFQEC*C*QAEDK	61	3833.91	59	96.7

Init. mass	*Seq. Pos.*	*Sequence*	*Theo. H/D*	*H/D Mass*	*Exp. H/D*	*% H/D*
927.50	87-95	SDTSSMSGR	22	944.78	17	77.3
1146.60	233-242	IEPELSDSEK	20	NV		
1249.61	335-344	AC*SSEVM*M*FR	22	NV		
1283.72	324-334	LLQEDQIALLK	22	NV		
1537.80	458-471	TLGNQNSEMCISLK	30	NV		
1576.76	251-263	ISPEQEELILIHR	26	1592.92	22	84.6
1673.82	135-149	ASGYHYNALTC*EGCK	31	NV		
1907.91	70-86	IWIPGHTIIASNHHL AK	29	1936.03	28	96.6
2513.97	324-344	LLQEDQIALLKAC*SS EVM*M*FR	41	NV		

In order to further increase the throughput of protein identification and to offer a flexible and powerful proteomic visualisation tool, we designed a highly automated method that can create a fully annotated 2-DE map [139]. This technology called "molecular scanner", combines parallel methods for protein digestion and electro transfer in a PMF approach to identify proteins. MALDI-MS analysis is conducted directly on the PVDF membranes by a scanning procedure. Using a set of dedicated tools this creates, analyses and visualises a proteome as a multidimensional image. This provides the technological basis for the development of a clinical molecular scanner, which could, for example, be adapted to medical diagnostics [140].

2.3.1. Double parallel digestion process

At the 1998 Sienna conference, our group presented a "parallel protein digestion during the electro blot" system [141, 139, 142]. During electro transfer, a membrane (Immobilon™ AV or IAV), containing covalently bound trypsin (IAV-trypsin), was present between the gel and the PVDF collecting surface. This results in tryptic cleavage of proteins during their migration to the PVDF membrane. In that respect, the transfer tension was adapted to reduce the migration speed of the proteins using an alternative square shape tension. The resulting effective tension was 3.5 V and this electric field had also the advantage in that it modified the protein orientation during each pulse. With this technique, the problem of low recovery of basic and high M_r polypeptide after electro blotting were still encountered. An improvement was achieved by operating a pre-digestion of the proteins in the gel prior to electro blotting. This combination called "Double Parallel Digestion" (DPD), led to greatly improved digestion of high molecular weight and basic proteins without losses of low M_r polypeptides. This method allowed successful identification by PMF of proteins differing over a wide pI and M_r range directly on the collecting PVDF membrane using MALDI-MS [141, 139].

The whole procedure is carried out as described in a recent Bienvenut *et al.* article [141]. Briefly, after SDS-PAGE protein separation, gels are soaked 3 times in de-ionised water for 5 minutes and then air dried at room temperature for 12 hours (for example overnight). For the first pre-digestion, gels are re-hydrated with 0.05 mg/ml trypsin in 10 mM Tris-HCl, pH 8.2 during 30 minutes at 35°C. Subsequently, the gels are transblotted onto PVDF membrane in a laboratory-made semidry apparatus at room temperature. In order to increase the migration time of the protein through the IAV membrane during the transfer (and thereby allowing more time for digestion to take place), an asymmetrical alternating voltage was used. A square waveform alternating voltage was selected: +12.5 V for 125 ms followed by -5 V for 125 ms, repetitively. The transblotting process is completed after 12-18 hours. To perform the digestion during the electro blotting, a double layer of IAV-trypsin membrane is inserted between the polyacrylamide gel (where the proteins are located) and the PVDF membrane (which acted as the collecting surface), to create a transblot-digestion sandwich. After the transfer procedure, the PVDF membrane is washed in de-ionised water for 5 minutes and stained if required.

2.3.2. ^{14}C quantitation of the transferred product and diffusion

The technique of Western blotting is widely used and a lot of investigations have been undertaken in order to quantify protein recovery on the collecting membrane [143, 144, 145, 146, 147]. One of the most common difficulties related to the description of the DPD transfer process is the estimation of the yield of proteins transferred from the gel onto the collecting PVDF membrane. One of the solutions was to use ^{14}C radiolabelled proteins [193].

^{14}C activity is an emission of β^- particles of low energy easily absorbed by the environment. Due to the thickness of the gel, it is not possible to obtain an accurate measurement of the β^- signal emitted by the gel separated proteins. Therefore, an absolute quantitation of proteins recovered on the collecting membrane is not possible. To overcome this problem, the signals acquired on the collecting membranes were compared to a reference obtained from a one dimensional electrophoresis (1-DE) gel of the ^{14}C labelled proteins. Protein recovery under different conditions was measured and the influence of the following parameters were evaluated:

- Effect of the buffer (heterogeneous CAPS and homogeneous ½ Towbin);
- Effect of the electric field used for the transfer: $1 mA/cm^2$ or square shape tension (SST);
- DPD versus standard transfer.

Comparison of the influence of the electric field on the protein recovery

The efficiency of the transfer using standard transblotting techniques or adapted SST (used during the DPD process) was tested without the digestion step, the main parameters are shown in Table 9.

Table 9. Buffers and electrical fields used during the experiment for the comparison of recovery of undigested proteins

Experiment	*1*	*2*
Buffer	Heterogeneous CAPS	Homogeneous 1/2 Towbin
Composition	10 mM CAPS buffered at pH 11	13 mM Tris, 100 mM glycine
Anodic (MeOH)	20%	12.5%
Cathodic (MeOH)	5%	12.5%
Electric field	$1 mA/cm^2$	SST

For each of these experiments, 2 lanes of 1-DE mini-gel were used. One μg of Bio-Rad M_r standard was loaded on the first lane of the gel and 50 nCi of ^{14}C radiolabelled proteins from Amersham-Pharmacia Biotech were loaded on the second lane. Proteins were separated using a standard technique [39]. At the end of electrophoresis, the gel was washed for 3 minutes in de-ionised water, 1 minute in

the 20% methanol buffer. PVDF membranes were equilibrated in the 5% methanol buffer for the experiment 1. In experiment 2, the same homogeneous 1/2 Towbin was used to equilibrate the gel and the PVDF membrane. After transfer, the membranes were washed rapidly with de-ionised water and air-dried.

PVDF membranes were scanned with a Phospho-Imager apparatus for the ^{14}C radiolabelled proteins and with an optical densitometer for Bio-Rad M_r standard stained with Amido Black. The volume of the spots were measured using the Melanie software [148].

The image obtained from ^{14}C labelled samples (Figure 4) showed 7 bands corresponding to myosin (MYSS), phosphorylase b (PHS2), albumin (BSA), ovalbumin (OVAL), carbonic anhydrase (CAH2) and lysozyme (LYC). The results are summarised in Table 10.

Table 10. Percentage increase in protein recovery using DPD type transblotting process compared to standard transfer; NP: protein not present in the sample; NV: protein not visible on the PVDF after staining*

	Transfer technique	
Proteins	Bio-Rad M_r standard with amido black stain (experiment 1)	14C labelled protein with autoradiography (experiment 2)
MYSS	NV*	110
BGAL	36	NP
PHS2	37	70
BSA	6	10
OVAL	21	14
CAH2	20	21
ITRA	21	NP
LYC	28	18
Average	24	23

Both experiments showed the positive effect on protein recovery when using the square shape tension during the electro transfer process. The average increase is equivalent for the Amido Black stained and ^{14}C labelled proteins. Nevertheless, this electric field is not acting identically on all proteins. The effect is less important for low M_r proteins (less than 20% for proteins smaller than 60 kDa) and the strongest effect is found for the high M_r proteins i.e. MYSS, BGAL, PHS2.

DPD quantification test

The staining intensities of a protein decrease progressively as a function of the extent of digestion[149, 150] and thus a comparison of the staining intensities of the

loss of material was the use of ^{14}C radiolabelled proteins. The DPD process was compared to the standard protein electro blotting to PVDF membrane (i.e. experiment 2 using SST, see Table 9). The quantitation was done as above with the 6 ^{14}C labelled proteins (MYSS, PHS2, BSA, OVAL, CAH2, and LYC). Digestion control was performed in parallel with the Bio-Rad M_r standards. The results obtained after 24 hours of exposition are shown in Figure 5.

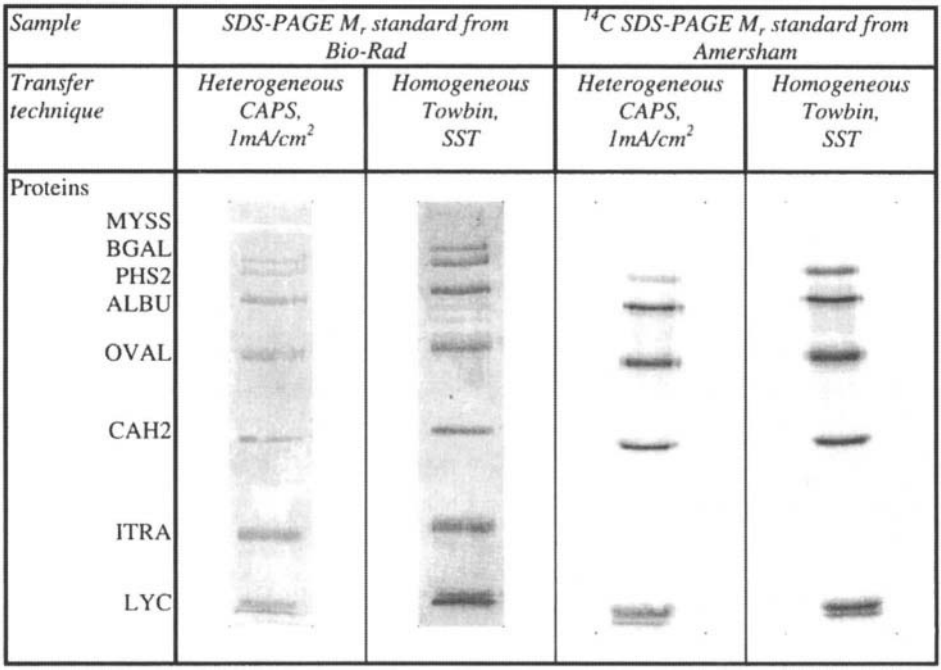


Figure 4. Amido Black stained proteins and auto-radiography of ^{14}C samples.

The binding activity of the PVDF membrane is mostly due to electrostatic interaction. Smaller peptides did not bind to the surface of the collecting membrane with a strong interaction. ^{14}C labelling of proteins is performed by modification of the γ-amino group of lysine [151]. Radioactivity decrease during the DPD technique could be due to loss of small peptides carrying C^{14} labelled lysine. This loss of a few small peptides does not seem to be a problem for protein identification. Interestingly, in the case of myosin and more generally for high M_r proteins, this technique allows a more efficient transfer of polypeptides to the collecting membrane than in the traditional transfer process.

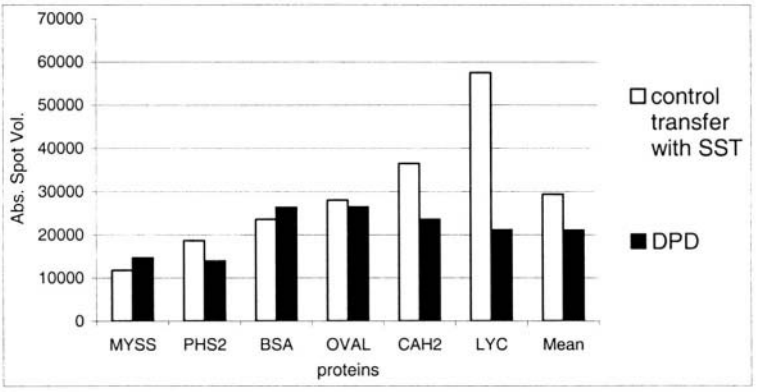


Figure 5. Absolute ^{14}C-signal intensity of the control transfer and DPD process.

3. PROTEIN IDENTIFICATION USING BIOINFORMATICS TOOLS

The field of protein identification has expanded over the last few years with the improvements of accuracy and sensitivity of MS instruments. At the same time, development of computer resources helped to speed up the analysis of the huge amounts of data generated by MS instruments, and to increase the number of nucleotide and protein sequence entries in specialised databases. Much progress has been made concerning high throughput facilities to prepare samples and run 2-DE gels. Furthermore, the automatic analysis from complete 2-DE gels up to the mass spectra data is already possible without human intervention [47, 139]. Many available PMF identification and post-identification software tools are able to assist with protein identification, but the final analysis of the results still requires human interpretation and validation.

Some bioinformatics software tools for proteomics combine data analysis, statistics and artificial intelligence methods to manage MS data, to identify proteins and to update databases. In this section, specific tools used to identify proteins are reviewed. They use lists of peptide mass values from MS or MS/MS as input, and they may also combine this information with amino acid sequence tag information or amino acid composition to enhance the identification of proteins. Figure 6 shows a simplified flow chart of sample preparation and MS data collection. It also shows the techniques and tools for protein identification described in this section.

Bioinformatics is also concerned by the huge amount of data generated experimentally by the wet-lab, as well as the data generated by its tools outputs. In this case, a laboratory management systems (LIMS) must be precisely designed for each specific need.

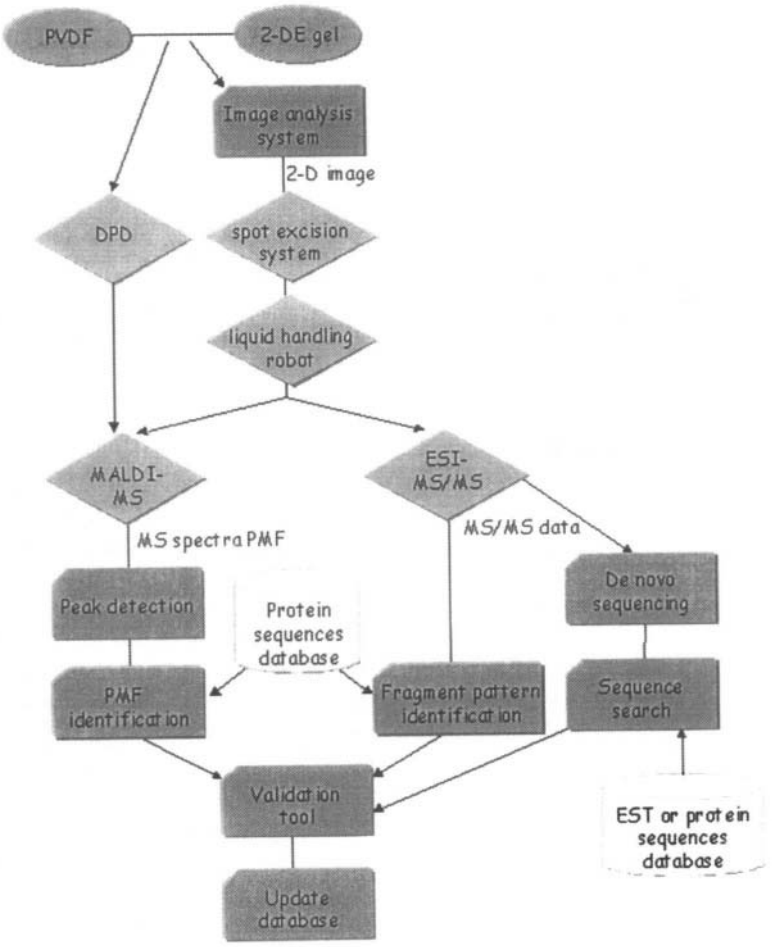


Figure 6. Possible schematic and simplified data flow for protein identification using mass spectrometry.

3.1. Protein identification by PMF tools using MS data

PMF, currently the most common method used to identify proteins in a high throughput environment, is based on the comparison of a list of experimental peptide masses with theoretical peptide masses. The experimental masses are generated from the MS measurement of an enzymatically digested protein sample. The theoretical masses are obtained from an *in silico* digestion of all sequences in a database. The goal is to find the protein(s) whose peptide masses show the best match with the experimental fingerprint. The method can be divided into 3 steps. The first step is peak detection, i.e. the selection of the most relevant masses for protein identification from the mass spectra. Frequently, only few experimental peptide

masses in the fingerprint match the theoretical masses, and it is therefore crucial to detect low-intensity but "important" peaks, while, at the same time, avoiding the selection of too many "non-important" peaks. The second step is the comparison of the selected experimental peptide mass values to all protein sequences in a database, which were theoretically cleaved by applying the cleavage rule corresponding to the enzyme used for the sample digestion. Finally, a similarity rule (score) should provide a measure of quality of fit of the matched values, in order to either automatically interpret the result and choose the best-matching protein, or to help the user to identify the correct protein.

In the comparison phase, apart from the experimental peptide masses and the proteinase used to digest the proteins, some optional attributes may be specified to reflect experimental conditions and to reduce the search space. These optional attributes may include information coming from the sample such as species of origin, M_r or pI of the whole protein with the accepted error range, possible chemical or artefactual modifications like carboxymethylation of cysteines or oxidation of methionines. Other parameters to be specified include the mass tolerance or the minimum number of matching peptides required for a protein to be suggested as a possible match. Providing a maximum of information available about the sample helps to decrease the number of candidate proteins, to reduce the probability of false positive matches, and thus to increase the confidence of the identification. However, one must be careful not to miss the correct protein either.

3.1.1 Peak detection

Peak detection is an important step in the identification process. Sometimes only a few experimental peptide masses in the fingerprint match the theoretical masses, and therefore the failure to detect a relevant peak can hinder the correct identification of a protein. However, if too many false peaks are considered, this may lead to erroneous database matches causing false identifications, as well as increasing search duration. Furthermore, it is important to precisely determine the peptide masses.

Algorithms that perform peak detection usually take into consideration the probable isotopic distribution when looking for the relevant monoisotopic masses. For example, Breen *et al.* [52] use a Poisson model to calculate the isotope distribution in order to select the monoisotopic peaks. These algorithms should also be able to separate overlapping isotopic patterns.

In some cases, peak detection software's delivered with the spectrometer hardware, designed to determine the monoisotopic masses, do not have the necessary flexibility. In our case, for example, the peak detection software had to be rewritten in order to be integrated into the automated high throughput identification pipeline. A genetic algorithm was proposed to optimise the thresholds needed for peak detection [53]. We have then shown the important correlation between peak detection thresholds and identification results.

3.1.2 Identification Tools

Several tools are available to identify proteins using PMF. They all compare peptide masses obtained from mass spectrometry experiments to the theoretical peptide masses obtained from a theoretical digestion of all sequences in a protein sequence database.

The programs generate a list of protein entries, ordered by a score that tries to reflect the fit between theoretical and experimental parameters. It is therefore evident that the order of suggested proteins in the result list is of paramount importance for a facilitated interpretation of the identification, in particular when manual intervention should be minimised. All programs compute scores for each hit; some of these scoring systems are very simple, while others use probabilistic methods to increase confidence in the matching protein. A list of PMF tools and their URLs is given in Table 11.

The simplest scoring method counts the number of peptide masses matched. This is applied by the PeptideSearch tool (http://www.mann.embl-heidelberg.de/Services/PeptideSearch/PeptideSearchIntro.html) which queries a non-redundant database (nrdb), as well as by our PeptIdent program [152] which searches the SWISS-PROT and TrEMBL databases [153]. When using tools based on this scoring approach, it is important to note that an upper boundary for the intact protein mass should be specified, since a score based on the number of matched peptides alone, clearly favours high molecular weight proteins. The MOWSE program [154] determines a score by considering the frequency of each peptide mass in the NCBInr database, process giving stronger weights to heavy peptides, as these peptide masses can be observed less frequently. This score also takes into account the presence of missed cleavage sites in matched peptides: the user can select to down-weight the contribution of partially cleaved peptide fragments to the score, by specifying a value for the so-called pFactor. The MS-Fit program[68] uses a similar scoring method, and can be used to search several databases, including NCBInr, GenPept, pdbEST, and SWISS-PROT.

The algorithm of the Mascot program [155] is based on the one used by MOWSE, but it introduces a probability-based score, which considers the matches as random events depending on the number of entries in the database. ProFound [157] calculates a probability for the identification of the correct protein, given by a bayesian formula, and uses the distance between experimental and theoretical masses obtained from the NCBInr database. A more recent version of this tool (see Table 11 for the URL) takes into consideration many more attributes and the same scoring method (personal communication). For the sake of clarity, it will be referenced in this document as ProFound-New. Finally, the MassSearch program [66] also determines an identification score based on the probability to randomly obtain a match of n experimental masses with n theoretical masses, given the interval of possible masses and the maximum allowed distance of masses accepted in this match. This process is repeated through epochs. At each epoch, a new mass is added, by increasing the allowed maximal distance, until a maximum probability is reached.

Table 11: Programs freely available on the Internet for protein identification

* OWL was last updated in May 1999 (release 31.4)

	Program Name	*Scoring Type*	*Search Database*	*Internet URL address*	*References*
Protein attribute (MS or MS/MS)	Mowse	Probabilistic models	OWL*	http://srs.hgmp.mcr.ac.uk/	Pappin *et al.*, 26
	Mascot	Mowse	OWL*/NCBInr	http://matrixscience.com/	Perkins *et al.*, 155
	MS-Fit/MS-Tag/MS-Seq	Mowse	SwissProt/Genepept/pdbEST/OWL*/NCBInr	http://www.prospector.uscsf.edu/	Clauser *et al.*, 68
	ProFound	Bayesian algorithm	NCBInr/ SwissProt	http://prowl.rockefeller.edu/cgi-bin/ProFound/	Zhang and Chait, 132
	PeptideSearch	Number of peptides matched	NCBInr	http://www.mann.embl-heidelberg.de/Services/PeptideSearch/PeptideSearchIntro.html	
	PeptIdent	Number of peptides matched	SwissProt/TrEMBL	http://www.expasy.ch/	Binz *et al.*, 139
	SmartIdent	Heuristic score based on learning method	SwissProt/TrEMBL	http://www.expasy.ch/	Gras et al., 51
	PepFrag		SwissProt/PIR/NR/dbEST and others	http://prowl.rockefeller.edu/PROWL/pepfragch.html	Fenyo *et al.*, 159
Other tools	SeqMS	Probabilistic models		http://www.protein.osaka-u.ac.jp/organic/SeqMS.html	F-Cossio *et al.*, 167
	MassXpert			http://frl.lptc.u-bordeaux.fr	
	GlycoMod			http://www.expasy.ch/tools/glycomod/	Cooper *et al.*, 174

All these algorithms use various attributes (in addition to the mass values) to limit the number of candidate proteins. However, they make little use of this information in their score calculation, since they use at most one or two of these attributes, such as the presence of missed cleavage sites or the mass distribution in the database. They represent only a small part of the parameters that could influence the quality of identification. We proposed a scoring scheme that considers about 30 attributes with their respective contributions to the score values [66]. As the importance of these contributions is difficult to estimate, this approach uses a learning algorithm: a genetic algorithm has been implemented to estimate the weight corresponding to each specific attribute. As a result, the discrimination rate of candidate proteins can be enhanced, i.e. the score allows to distinguish between false positive and correct matches. This method, which is implemented in the SmartIdent tool, is also robust against mass calibration errors, since it uses a linear regression method to qualify the global goodness of the matches between the experimental masses and the theoretical ones.

Table 12 shows a comparison of the results of some of the available PMF programs when analysing a very difficult mass spectra. The query was made for a list of 60 masses measured on a Voyager Elite MALDI-TOF MS. The sample corresponds to protein G3PC_ARATH (SWISS-PROT P25858) obtained after 2-DE separation and tryptic digestion as described by Bienvenut *et al.* [141]. In this comparison, the parameter values used in all identification programs were identical, whenever possible, and they are detailed in the table legend. In the case that the parameters were not comparable, the default parameters were used. The difficulties highlighted by the use of such different programs to identify the correct protein are very interesting. SmartIdent, MS-Fit, ProFound and ProFound-New with SWISS-PROT database retrieved the correct protein entry as the first hit. PeptIdent with SWISS-PROT database ranked this protein in 2^{nd}, with the same score as the first hit, since they both have the same number of peptide masses matched. PeptIdent with SWISS-PROT and TrEMBL database, Mascot and ProFound-New with NCBInr database have assigned ranks as high as 8 to this protein.

When querying the NCBInr database, the identification ranks are completely different depending on which PMF tool is used. For the example shown in Table 12, MS-Fit and ProFound ranked G3PC_ARATH at the top level while Mascot and ProFound-New ranked this protein with a lower score. The MS Fingerprinting Spectrum Analysis Tool from Starlab, Gent (http://genesis.rug.ac.be/~nikos/) was also tested, even though it's scoring method was not described in this document (data not published, personal communication). It is also apparent that the size of the database has a strong influence on the PMF result. For example, on the SmartIdent output, utilisation of SWISS-PROT alone or SWISS-PROT/TrEMBL as target databases modified the level of discrimination of the result. This problem is mainly related to the database size (88757 entries in SWISS-PROT Release 39.7 and 300152 entries in TrEMBL Release 14.17). A larger database increases the probability to generate a false match.

In the outputs of the software's that ranked the protein correctly, the score values and the discrimination values gave good hints about the quality of the match, particularly when comparing the first hit against the second one. One must consider

that this mass fingerprint was very difficult to analyse and was specially chosen to outline the gambling aspects of PMF analysis.

Table 12: Identification of G3PC_ARATH (GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE,CYTOSOLIC), Species Arabidopsis thaliana, SWISS-PROT (SP) entry P25858 using different PMF tools. The restricting parameters used were Arabidopsis thaliana for the species, a minimum number of 4 matched masses, a maximal tolerance for masses of 60 ppm, at most one missed cleavage of tryptic peptides allowed, and the modifications accepted were oxidised methionines and cysteines treated with iodoacetamide to form carboxyamidomethyl cysteines. The databases queried by each program are listed in the right column. In the score column, the first value is the score of the first candidate protein, followed by either the score of the second candidate protein (if the first one is the correct one) or the score of the correct protein

Program Name	*Rank*	*Score*	*Discrimination value/rule*	*Database*
SmartIdent	1	130.41 (16.51)	0.89	SP
	1	130.41 (80.40)	0.62	SP/TrEMBL
PeptIdent	2	0.08 (0.08)	-	SP
	3	0.15 (0.08)	-	SP/TrEMBL
Mascot	8	41 (29)	If score > 58, then significant	NCBInr
MS-Fit	1	3890 (156)	-	SP
	1	634 (633)	-	NCBInr
ProFound	1	-	0.36 (0.29)	NCBInr
ProFound-New	2	-	1.0 (0.00019)	NCBInr
	1	-	0.99 (0.01)	SP
PeptideSearch	9	-	-	SP/TrEMBL
Startlab (Gent)	1	6 (6)	-	nrdb

Table 13 presents the analysis of PMF results for a mixture of two proteins. PMF of the protein PON2_HUMAN (SERUM PARAOXONASE / ARYLESTERASE 2, Species *Homo sapiens*, SWISS-PROT accession number Q15165) and the marker GT26_SCHJA (GLUTATHIONE S-TRANSFERASE 26 KDA, Species *Schistosoma japonicum*, SWISS-PROT accession number P08515) were obtained after 1-DE separation and tryptic digestion as described by Bienvenut *et al.* [141]. All programs have found the marker as the first hit, and PON2_HUMAN protein as the second hit. In the cases where PON2_HUMAN was not the second hit, the preceding hits on the list were variants of the marker. ProFound offers the possibility to decide if the mass values correspond to one single protein or to a mixture of up to four proteins. Selecting one single protein or a mixture of two proteins does not change the results, i.e. PON2_HUMAN and GT26_SCHJA are always at the top of the hit list with high scores. In fact, these results show that the performances of PMF tools highly depend on the search databases.

Table 13: Query for 60 mass values of a MALDI-TOF MS spectrum from a mixture of PON2_HUMAN (SERUM PARAOXONASE/ARYLESTERASE 2, Species Homo sapiens, SWISS-PROT accession number Q15165) and the marker GT26_SCHJA (GLUTATHIONE S-TRANSFERASE 26 KDA, Species Schistosoma japonicum, SWISS-PROT accession number P08515). The query was made for the all available species, the minimum number of matched masses was 4, the maximal tolerance for masses was 40 ppm, at most one missed cleavage for tryptic peptides was allowed, and the modifications accepted were artefactual modification of cysteines with acrylamide and oxidised methionines. M_r values were delimited between 20 and 40 kDa. The databases queried by each program are listed in the right column

	GT26_SCHJA		PON2_HUMAN		Database
Program Name	Rank	Number of peptides matched	Rank	Number of peptides matched	
SmartIdent	1	15	6	10	SP/TrEMBL
PeptIdent	1	15	7	10	SP/TrEMBL
Mascot	1	16	38	8	NCBInr
MS-Fit	1	15	2	10	NCBInr
ProFound	1	15	2	5	NCBInr
PeptideSearch	1	11	25	7	SP/TrEMBL TrEMBLnew
Startlab (Gent)	1	13	32	9	nrdb

All the available tools have their own characteristics, and each of them has its own strength and weakness. It is therefore not surprising that they can produce quite different results for the same sets of peptide masses, particularly for "difficult" query spectra such as the case of protein G3PC_ARATH. With the aim of giving users the opportunity to take advantage of the features of a large number of tools whilst having to fill in only one single submission form, the CombSearch tool (http://www.expasy.ch/tools/CombSearch/) was designed. It simultaneously submits the specified input data to several protein identification tools available on the internet, and tries to assist in integration of the results.

3.2 MS/MS Ions Search

Peptide mass fingerprinting characterises each peptide by only one attribute, the peptide mass value. By itself, a single mass value does not reveal much about the peptide or the protein sequence, however other protein attributes such as ions from internal sequences of peptides obtained from successive MS fragmentation may give better hints for protein identification. Algorithms similar to those of PMF are also used for MS/MS ions search. In general, all proteins contained in a database are digested *in silico* to find parent peaks. These theoretical parent peptides are then fragmented *in silico,* and the experimental MS/MS pattern is compared to the theoretical patterns [157, 158]. The discrimination score is given by correlating theoretical and experimental fragments.

Some of the scoring methods of peptide mass fingerprinting tools are used in the analysis of MS/MS mass spectra of peptides. PeptideSearch, for example, counts the number of experimental and theoretical fragments matched, whereas Mascot uses a probabilistic score (personal communication). The Sequest program [157] is a blend of two approaches which runs in two sequential steps. The first step counts the number of matches among all ion fractions. The second step takes a certain number of the best matches and builds a theoretical spectrum where the peak intensities depend on the ion types. The score measure is then given by correlating theoretical and experimental spectra through their Fourier transformation and choosing the best correspondence in this space. In a more recent version, called Turbo Sequest (http://www.thermoquest.com/turbosequest.html), the computation time required for the first step was reduced by the use of an indexed database. In this case, the indexes are based on the peptide masses.

Most of these algorithms are well adapted to fragment peptides by CID, meanwhile the tool available at Starlab, Gent (http://genesis.rug.ac.be/~nikos/) is adapted to fragment ions by PSD. Even though this tool was presented publicly in Siena 2000 [104], its scoring method has not yet been published.

Other possibilities to improve the results of MS/MS data comparison are to combine this information with peptide sequence tags or the amino acid composition. This information is taken into account in programs such as MS-Tag, MS-Seq, PepFrag [158] and PeptideSearch.

The identification of proteins with MS/MS is a powerful technique specially for the identification of protein mixtures [160]. There is also the possibility to search in expressed sequence tags (EST) databases [161]. However, it is also important to note that many more modifications are possible when fragmenting the peptides, which result in exponential combinatorial possibilities of search.

3.3 De novo sequencing

Very often, no exact database match can be found even with high quality MS/MS mass spectra. It depends directly on the completeness and accuracy of the database searched, i.e. whether the genome is complete or incomplete, and on the quality of the transcripted EST sequences. These problems raise the question whether it is a novel protein, a known protein with a post-translational modification or if the failure to produce a database match due to inter-species variation, database sequence errors, or unexpected proteolytic cleavages. To address this problem, *de novo* interpretation of MS/MS data is an alternative where the amino acid sequence of peptides is derived by interpreting the mass differences between the generated MS/MS fragment ions sequence.

The automatic or visual interpretations of MS/MS data require considerable efforts. *De novo* peptide sequencing algorithms generate results that may be ambiguous since the analysis of MS/MS ions is not a simple task. Important experimental problems such as noisy and incomplete data, or ion types dependent on the ionisation method, are some of the problems to be addressed.

Although no tools, freely available on the Internet, exist that integrate automatically protein identification with *de novo* sequencing algorithms, some

isolated programs deduce peptide sequences from a list of ion fragmentation masses. Their algorithms are mainly divided in two different approaches. The first described global approach generated all possible fragments of amino acids, compared them to theoretical fragments and used to keep the best matches [162]. Prefix pruning algorithms were also applied to reduce the combinatorial explosion, since the number of sequences grow exponentially with the length of peptides [163, 164]. This approach restricts the search space to sequences that best match the experimental spectra according to their prefix. The inconvenience of this method lies in the fact that regions of a sequence that are under-represented by fragment ions may be discarded before any analysis.

The second paradigm of a local search provides more efficient results and is based on graph theory. The peaks in an experimental spectrum are transformed into vertices in a spectrum graph, where the edges correspond to differences of masses between two ions. According to experimental conditions, these mass gaps can result from some amino acids, single amino acids, fragments of amino acids or modifications [165, 166, 167]. The solution is given by searching for the path in the resulting acyclic graph which has the best score. Scores may be calculated based on probabilistic models, implemented in the so-called Sherenga algorithm [168] and in the SeqMS tool [167], or on a combination of ion intensities and cross-correlation [165].

For Zhang and McElvain [169], the peptide sequences are obtained by reading the intersection spectrum of a MS/MS daughter ion and their granddaughter ions MS^3 (a third stage of mass spectrometry). This intersection spectrum represents common peaks to both MS^2 and MS^3 and are obtained through an arithmetic mean of the ion intensities. This algorithm also uses cross-correlation and ion intensities to calculate the similarity score of the different spectra.

Other innovative algorithms for *de novo* sequencing have been developed based on learning methods from artificial intelligence. Stranz *et al.* [170], instead of using the usual graph theory, proposed an adapted genetic algorithm to optimise the search for possible combinations of amino acid masses. The Sherenga algorithm [168] automatically learns fragment ion types and intensity thresholds from a collection of test spectra. This information is then used to help *de novo* interpretation of peptides. Scarberry *et al.* [171] trained artificial neural networks to classify observed ion fragments into specific ion types (y, z, b, …) before deriving the sequence spectra.

3.4 Other tools related to protein identification

A range of proteomics tools are available that allow us to go beyond simple protein identification. Most proteins from higher eukaryotes undergo co- and/or post-translational modifications. These modifications can involve either a cleavage process (thus eliminating signal sequences, transit or pro- peptides and initiator methionines) or the addition or removal of many different simple chemical groups (e.g. hydroxyl, carboxyl, acetyl, methyl, phosphoryl, etc.), as well as the addition of more complex molecules, such as sugars and lipids. In these cases, the *in silico* calculation of peptide masses using non-annotated databases will not match the masses obtained experimentally any longer. Needless to say that these modifications

are important assets of proteins and therefore need to be characterised in order to describe the mature protein. A comprehensive tool for high throughput mass spectrometric discovery of protein post-translational modifications is the FindMod tool [172] available on the ExPASy server (http://www.expasy.ch/tools/findmod/). This tool considers some 30 post-translational modifications, applying many different rules derived from documented post-translational modifications in SWISS-PROT [153] and from the PROSITE protein family and domain database [173]. FindMod can also suggest possible single amino acid substitutions. Such substitutions can occur in proteins translated from polymorphic genes.

A similar tool that assists calculation and prediction of possible post-translational protein modifications is GlycoMod [174]. This tool deals only with protein glycosylation, probably the most common and complex type of protein modification. GlycoMod is available on the ExPASy server (http://www.expasy.ch/tools/glycomod/) and does not only allow for computing possible monosaccharide compositions corresponding to the mass of a glycopeptide, but also allows inclusion of a range of options such as oligosaccharide release or derivatisation strategies in the calculation.

Peptide masses that do not immediately match theoretical masses in the identification process, can not only be the result of a post-translational modification, but may arise during the processing of the sample. A tool that can identify possible peptides that have resulted from non-specific chemical or enzymatic cleavage of proteins is the FindPept tool (http://www.expasy.ch/tools/).

At this point we would also like to mention a versatile tool to model mass spectra and mass fragmentation spectra *in silico*. A range of proteomics tools is combined in MassXpert, a free application by F. Rusconi, which can be downloaded from http://frl.lptc.u-bordeaux.fr). Supported features include cleavage of a protein by different enzymes, fragmentation of peptides or small proteins, an m/z calculator for a given protein mass and many more.

3.5. Data storage and treatment with LIMS

The combination of 2-DE gels with mass spectrometry results in huge amounts of data. It is essential to manage this data in a centralised way in order to simplify its analysis and database updates. The bioinformatics analysis of experimental data also produce results and data, that must be managed and available at all times. A LIMS (Laboratory Information Management System) is a software application that uses a relational database to assemble heterogeneous data, such as gel images, mass spectra, samples, experiments, and related documents, and also provides the tools to allow such data to be entered, tracked and reported [175].

A LIMS has or should have the following characteristics:

1. Be an instrument management that allows centralised storage of maintenance and calibration records.
2. Be a data management that enables the complete laboratory environment to be mapped onto database. This allows organising information about personnel, instruments, analytical methods, work procedures and costs.

3. Have a wide range of validation techniques that ensures the integrity of data.
4. Be a sample management, which provides a variety of techniques for registration, processing, authorisation and archiving of routine and non-routine samples, standards and reference materials, as well as commonly used test sequences.
5. Have resource management capabilities that include instrument backlog reporting, as well as personnel time management. In addition costs associated with analysis may be calculated and invoiced to client accounts.
6. Have a communication management that ensures that important information reaches decision-makers with minimal delay. Access to data within LIMS may be achieved through a wide variety of mechanisms including the industry standard SQL. Information can be communicated using all common networks and interfaces.
7. Have a quality management that is particularly important in regulated environments and can be achieved through audit trail and validation facilities. Quality control is enhanced through specification libraries and action triggers with graphical data interpretation.
8. Have a security configuration that allows standard functionality to be used to create a secure system. This includes setting up passwords, authority levels and menus for each user.

A LIMS should provide a great flexibility and variety of functions without creating a system that requires expensive hardware platforms or one that is difficult to change, e.g. small changes require major system rewrites. The best LIMS are those that have simple or straightforward software architectures and run on inexpensive platforms or networks. Invariably, the keys to their success are flexibility, adaptability, ease of evolution and support, and most importantly overall system speed. The speed issue is very critical, as personnel will not use something that is slow or awkward.

An example of simplified client/server architecture of a LIMS is given in Figure 7. To allow clients to access the LIMS the system is centralised on a Web server. When connecting to the server, users receive, via the network, the forms corresponding to their request. Once completed, these forms are sent back to the server which processes the information accordingly. Useful information is stocked in the LIMS database. The software used in each stage may be, for example:

Server: Java Web Server, IIS, Apache etc,
Database: Oracle, Access, Mysql, etc,
Network: Internet,
Client: PC, SUN, etc,
Forms: Created with servlets (Java) or CGI and Perl.

LIMS are very specialised software's which require great financial and time investments. This may explain that freely available solutions have not been reported until today, and one of the commercial versions adapted to proteomics is BioLIMS from PE Informatics.

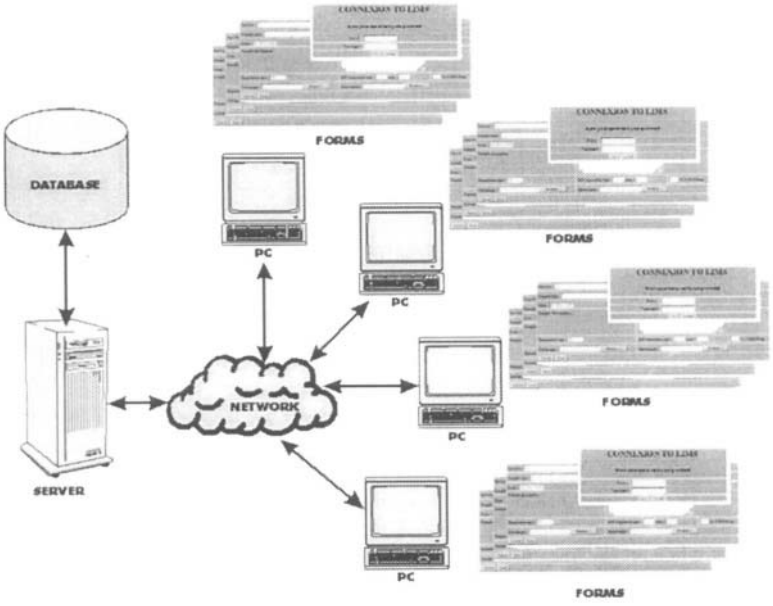


Figure 7. The architecture of a simplified client/server implementation of LIMS

3.6. Concluding remarks

One of the main objectives of research in proteomics is the automation of all procedures from sample acquisition to protein identification. The role of bioinformatics is to treat all available data to obtain unique identification of proteins without human intervention. The recent generation of mass spectrometers are producing more accurate data, with precise mass values using MS^n where n are successive stages of mass spectrometry [176, 177].

The complexity of the PMF must also be explored to optimise the identification tools by means of theoretical models that predict PMF spectra. Currently, most of the tools use a minimum of information to identify proteins. As described in this section, experimental information required for protein identification is limited to a list of mass values which is compared to theoretical fragments. Rules used to produce such peptides do not reflect spectra complexity due to various factors. Mass spectra samples contain peptides that are due to specific endoproteolytic cleavage of the target protein but also unspecific cleavage [82, 178] as well as artifactual peptide modifications. The theoretical rules used for *in silico* protein digestion are often much simpler that the real ones [82]. It is also difficult to prevent sample contamination from different kinds of contaminants, e.g. keratin and endoprotease autolysis products or disturbing agents i.e. SDS or non-volatile salts that could also

affected the signal. The chemical and physical properties involved in the mechanisms of ions formation are particularly complex [179, 180, 181] and as a consequence they are impossible to be modelled.

Preliminary studies towards the development of efficient spectra modelling tools have been published. These include, for example, the analysis of the distribution of peptide masses generated by *in silico* digestion tools from protein sequence databases [182, 154] and the influence of a sequence on spectra intensity [143].

Bioinformatics tools should be easily adaptable to integrate all the new data generated by their analyses. We should envisage that one day an automatic system would integrate the experimental data (such as a LIMS), all kinds of mass spectrum values (MS, MS/MS) and algorithms (PMF, ions search, *de novo* sequencing) to produce reliable and automatic protein identifications.

4. BIOINFORMATICS TOOLS FOR THE MOLECULAR SCANNER

As described previously, 2-DE gels are a method of choice to separate a large number of proteins with high resolution. If properly stained, a gel provides a two-dimensional graphical representation of a proteome.

Peptide mass fingerprinting is at present one of the most effective and rapid methods of identifying proteins excised from a 2-DE gel. In order to apply this method to a large number of spots various approaches have been described [119, 47]. The molecular scanner approach [139] combines parallel methods for protein digestion and electro transfer [141] with peptide mass fingerprinting methods. Two scanning experiments will be discussed : In the first one, a sample of 1mg *E. coli* was separated with a mini 2-DE gel and the rectangle excised from the collecting PVDF membrane was scanned on a 48x32 grid with a sampling distance of 0.25mm in both directions [141]. For the second experiment, human plasma was used and the membrane scanned was scanned on a 80x16 grid with a sampling distance of 0.25mm in horizontal direction and 0.5mm in vertical direction, respectively [139].

The software tools used to analyse the spectra, to perform the identifications and to visualise the results are described in this section.

4.1 Peak detection and spectrum intensity images

By means of a home-made sample and data pipelining software, the user can launch an analysis of these data. Firstly, the peptide peaks have to be detected in all spectra. Secondly, the user needs to create a virtual image that shows the presence of proteins on the membrane. The height of a peptide peak depends on many parameters and does not give good quantitative information on the concentration of peptides [80]. But, if we sum up the heights of all the peaks detected in a spectrum, this gives valuable qualitative information on the presence of proteins that can be obtained. By repeating this for every sampling point of the scan we can obtain an image representing the intensity of the spectra as depicted in Figure 8C and 8D.

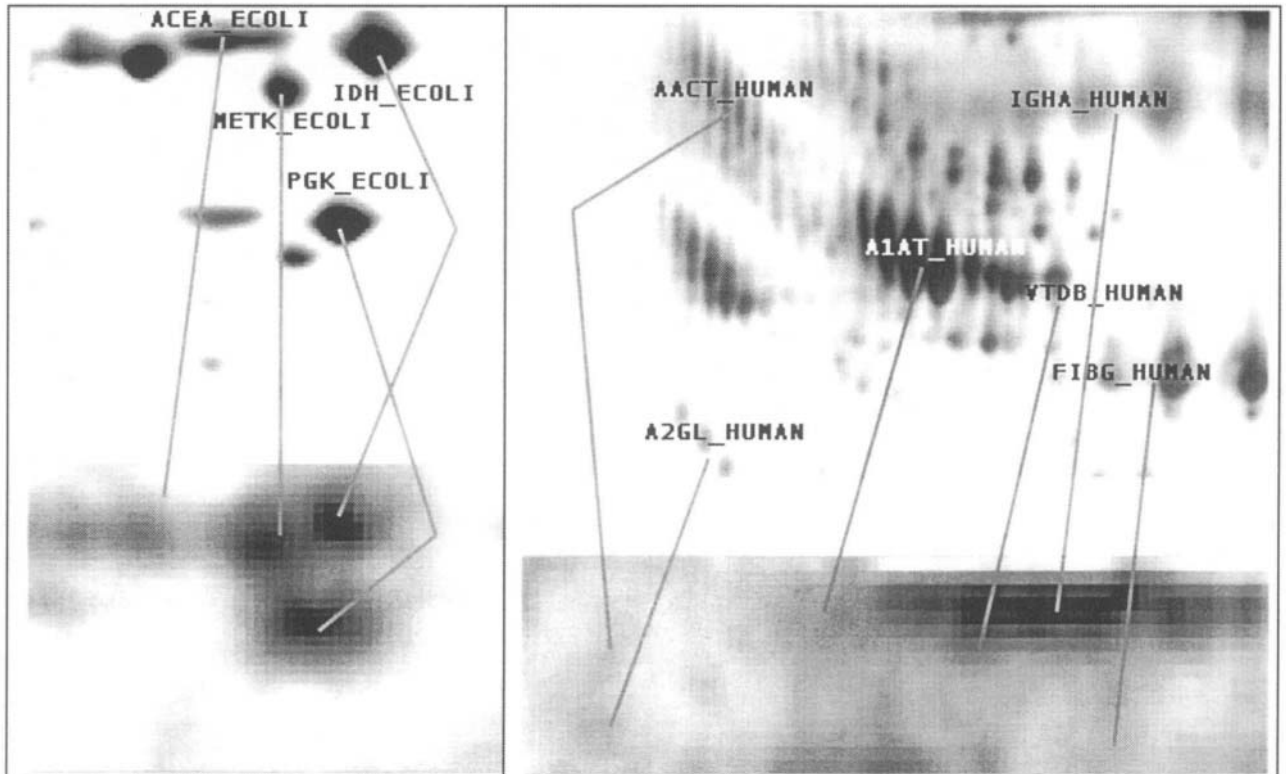


Figure 8. E. coli (A,C) and human plasma (B,D) reference gels (A,B) and spectrum intensities (C,D) (throughout this section high intensities are black and low ones are white). (A) Portion (pI range 5.1-5.2, M_r 35'000-45'000 Dalton) of the SWISS-2DPAGE reference gel (http://www.expasy.ch/cgi-bin/map2/def?ECOLI4.5-5.5 with annotated spots. (B) Portion (pI range 4.2-5.6, M_r 43'000–65'000 Dalton) of the SWISS-2DPAGE reference gel (http://www.expasy.ch/cgi-bin/map2/def?PLASMA_HUMAN) with some annotated spots. (C) Total intensity of peaks in E. coli and (D) in human plasma scan, where grey lines indicate spot identities. In order to obtain a better correspondence with the master gels, the images (C)-(D) were smoothed.

Figure 8 shows that there is a good correspondence between the SWISS-2DPAGE [183] and intensity images. Since the SWISS-2DPAGE reference gels were run with an acrylamide concentration gradient in the second (M_r) dimension and the mini 2-DE gels were not, the M_r scales of the reference gels and the intensity images are different. While there was good correspondence for the *E. coli* scan, the correspondence was less obvious for the human plasma scan because of the proximity of the immunoglobulin and albumin spots whose peptides were abundant and disturbed the ionisation of peptides from other proteins.

4.2 Protein identification

The purpose of the molecular scanner is for the identification of proteins that were separated with a 2-DE gel. Therefore, for each scan point, the lists of peptide masses are submitted to the peptide mass fingerprint identification program SmartIdent [51], which searches the protein sequence database SWISS-PROT and returns a list of matching proteins and their score.

The resolution power of a 2-DE gel is limited, and therefore several proteins may be found in the same position in a gel [184]. In MALDI MS, the presence of one peptide can attenuate the signal of another and some peptides are difficult to detect [80], resulting in a limited number of peptides per protein expressed in the spectrum. For the *E. coli* and human plasma scan the protein concentration on the PVDF membrane was low and we had to set the minimal number of matching peptide masses to 3 in order to identify some weekly expressed proteins. The mass tolerance was set to a value as high as 0.6 Dalton due to calibration errors and one missed cleavage was accepted.

Peak lists and identification results are then written into a text file. The 2-DE gel analysis tool Melanie [148] can read this file and allows loading the molecular scanner data for visualisation. It allows selecting scan points and viewing the identification results and spectra of these points.

Figure 9 shows a Melanie image of the human plasma scan, where two scan points were selected. In the point on the right side, immunoglobulin-α-1-constant-region protein (ALC1_HUMAN) was detected with the best score and good spatial correlation, i.e. the same protein was detected at least two times in the eight surrounding points. The following proteins in the scoring list have a significantly lower score and some of them are isolated identifications (indicated by an asterisk). α-1-antitrypsin (A1AT_HUMAN) matched with the best score in the other selected point, whereas the next protein (STK2_HUMAN) has a similar score. Two hypothesis could explain this observation: Either both proteins are present or there is a false identification. This problem must be carefully investigated and rules must be defined that decide whether the identification is correct or false.

4.2.1 Validation of identifications

In PMF, false identifications occur if a protein matches by chance some peptide masses detected in a spectrum. These peptide masses might stem from other

proteins, from impurities or matrix clusters or might be erroneously detected peaks. The more selective the parameters for the identification program are, the lower is the chance for a false match, but the higher is the chance that a true match is missed [185].

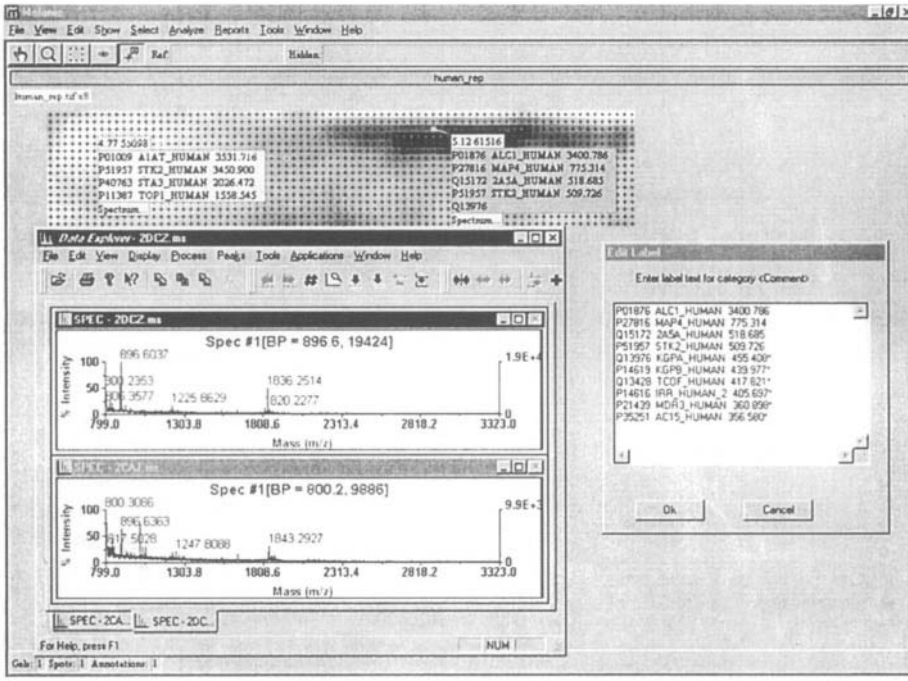


Figure 9. Identification data of the human plasma scan. The data were imported into Melanie and two scan points were selected. pI and M_r were calculated from the position of a point on the gel and are shown in the first field. The second field contains a list of the matching proteins (SWISS-PROT accession number and SWISS-PROT name) and their score. Clicking on this field displays the whole list (an asterisk means that the protein was not found a third time in a 3x3 neighbourhood of the point). Clicking on the third field renders the spectrum attached to the scan point with DataExplorer, which allows verifying peak detection and gives information about the intensity of the peaks.

The values chosen for the two scans were not very selective because the aim was to detect weekly expressed proteins. We therefore had to deal with a large number of false matches. To discard the bulk of them, two selections were applied. In order to avoid isolated matches, we applied a cellular automaton [186] to the list of matching proteins. A match was discarded if it was not found again in at least two of the eight neighbouring sites and this process was repeated until a stable configuration was reached. Then a threshold for the median score, i.e. the median of the scores of all sites where the protein was found, had to be set. Figure 10A depicts the distribution of the median protein score for the *E. coli* scan. In Figure 10B the peptide mass

fingerprints were compared with the mouse proteins in SWISS-PROT. It was assumed that the mouse proteins in the scanned p*I* and molecular weight window bear little resemblance with the *E. coli* proteins and thus provided a statistic for false matches. 95% of these false matches had an average score lower than 340, and this value was taken as the threshold.

Table 14. Protein identifications examined due to the criteria described in the text

SWISS-PROT Name	*Median Score*	*Status*	*Eliminating Criterion*
6PGD_ECOLI	4620.96	Ok	
IDH_ECOLI	4533.88	Ok	
METK_ECOLI	1860.03	Ok	
ALDA_ECOLI	1307.66	Ok	
PGK_ECOLI	1104.28	Ok	
DHPS_ECOLI	606.52	False	1,2
EUTB_ECOLI	508.55	False	2
FIXC_ECOLI	497.26	False	3
YAGE_ECOLI	475.95	False	2
YBHE_ECOLI	441.67	Ok	
ACEA_ECOLI	433.07	Ok	
6PG9_ECOLI	430.30	False	2
HSLU_ECOLI	354.67	False	1,2
ATOC_ECOLI	349.61	False	1,2

Table 14 shows the resulting list of proteins, which may still contain false matches. In order to purge them from this list, a detailed analysis of every protein was performed using the following criteria:

1. Spot shape
2. Matches with matrix cluster and impurity peaks
3. Reuse of matching peptide masses

All these criteria are rules of thumb and the user has to check them using visualisation tools. The first one serves to eliminate identifications that are poorly localised and therefore do not look like a spot. The second tries to eliminate matches with matrix cluster [187] and impurity peaks. For the third criterion it is assumed that every peptide mass belongs to one matching protein. If a protein matches with a lower score and reuses two or more peptide masses of proteins that matched with a significantly higher score, these masses are discarded. If these masses are indispensable for the match of the former protein, that protein will be eliminated from the list. A detailed discussion of these criteria will be given in Müller *et. al.* [188].

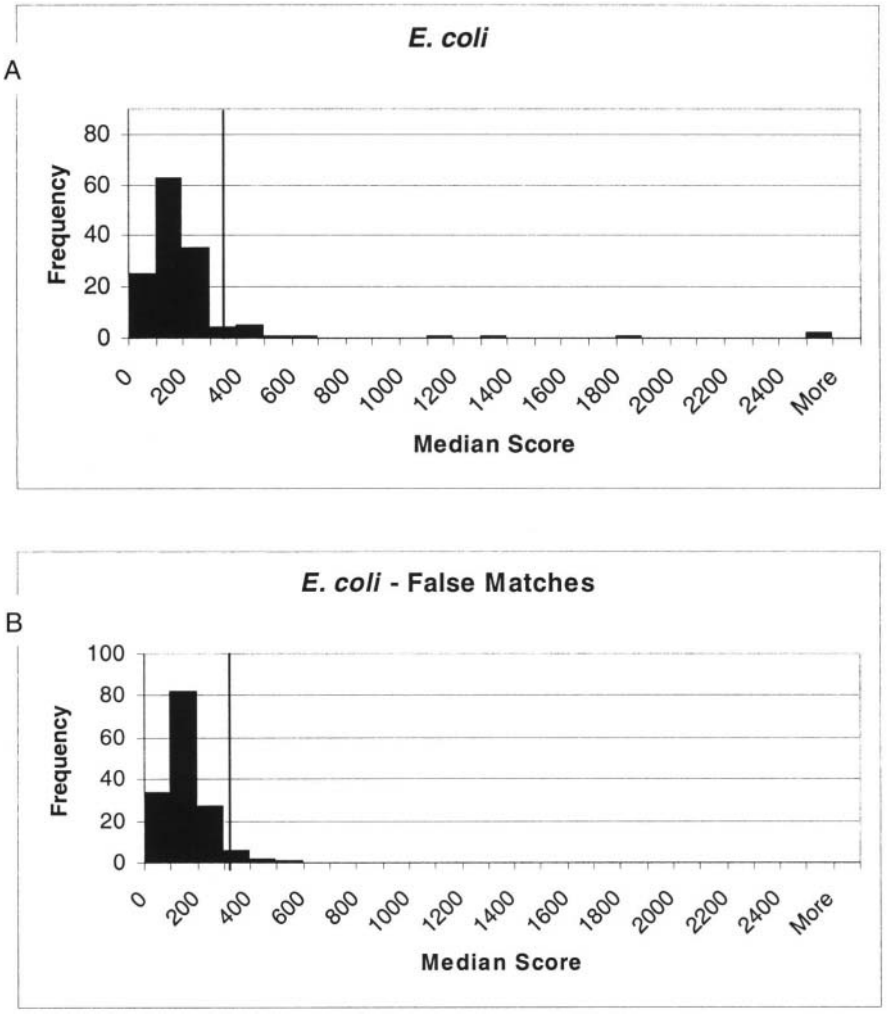


Figure 10. Histograms of the median score for the E. coli scan. (A) shows the median score resulting from matches of all E. coli proteins in the SWISS-PROT database, whereas (B) shows the average score of matches of all mouse proteins in SWISS-PROT. The matches in (B) provide a statistic for false matches, since the E. coli and mouse databases have a similar size (4602 and 4066 entries, respectively). The vertical lines mark the 95% confidence threshold.

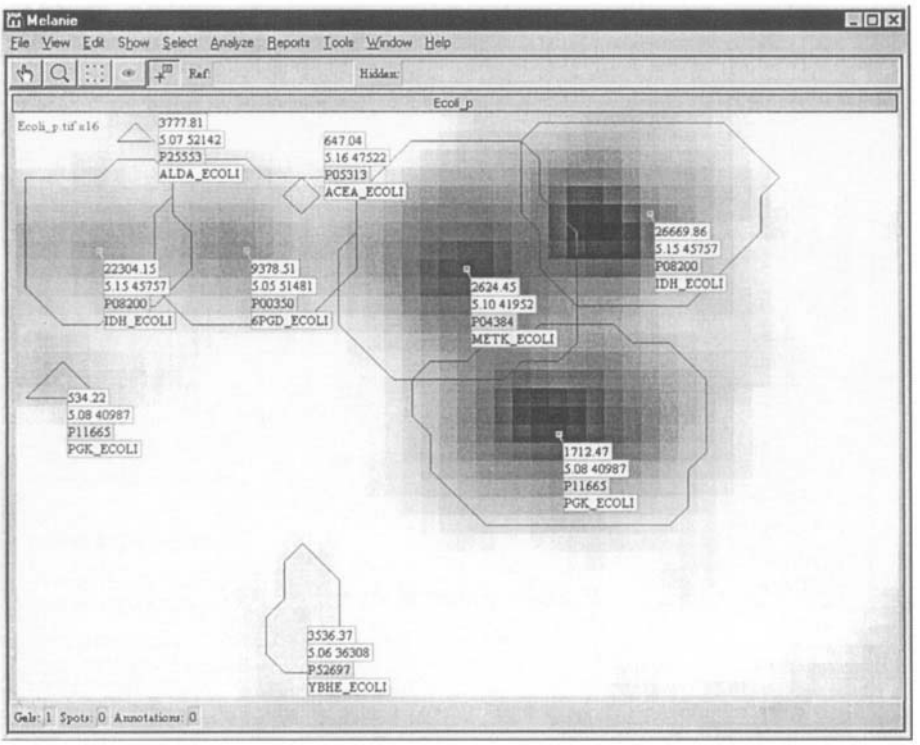


Figure 11. Spots detected in the E. coli scan. The spots were calculated using a dendrogram algorithm (Saporta, 189) and the contours enclose 50% of the points, where the protein was detected. The proteins PKG_ECOLI and IDH_ECOLI are split into two spots, which might correspond to modifications of these proteins.

An important point in criteria 1) and 2) is that they take account of spatial correlation and distribution of the data. This allows an improvement in the results to a much greater extent than if only localised information was available. We believe that this is one of the strongest features of the molecular scanner.

Figure 11 shows the proteins that fulfil all the above criteria. IDH_ECOLI and PGK_ECOLI were found in two spots. ACEA_ECOLI and ALDA_ECOLI were only weakly expressed and formed tiny spots. ALDA_ECOLI, 6PGD_ECOLI and YBHE_ECOLI were not identified on the master gel.

Apart from identifying spots automatically, a user might be interested in certain features of the data. An important question is whether a protein is in a modified form or not. The user can specify the mass change of this modification and the name of the protein under investigation and then draw a map where the modified peptides are found.

Figure 12 summarises the steps needed to analyse a scan. All steps apart from the last one can be fully automated. The automation of criteria 1) and 2) is a more difficult task and we are devising algorithms that can solve this problem.

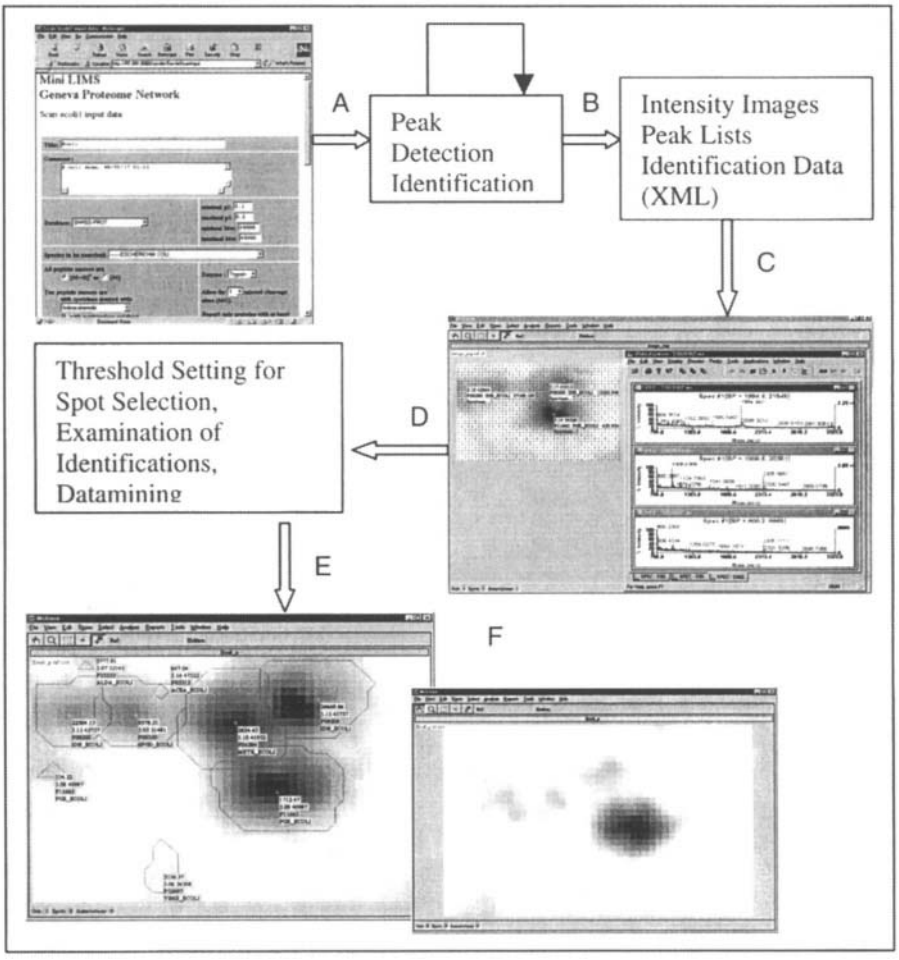


Figure 12 (A) By means of the LIMS, the user defines all the parameters needed for peak detection and identification (SmartIdent) and launches the data analysis program. For each scan point peptidic peaks are detected and the peak list is submitted to SmartIdent. (B) The results are written into TIF- and XML-files. (C) Melanie is used to visualise them. The user can examine identifications and (D) set thresholds for the spot detection. (E) All the proteins that pass these thresholds are selected and the contours of the regions where they were detected are drawn. (F) The intensity of peaks within 2025± 0.7Da is shown. The region where these masses are detected corresponds well to the PGK_ECOLI spot on the right side. Since these masses do not appear in the list of standard peptide masses of PGK_ECOLI, they might correspond to a modified peptide.

4.3 Concluding remarks

The molecular scanner is a powerful tool to analyse an entire proteome based on 2-D gel electrophoresis and PMF. For every scan point it yields a list of peptide masses and, after searching a peptide sequence database, a list of matching proteins. The two dimensional structure of the data allows easy visualisation and comparison of samples and the spatial correlation can be effectively used to enhance the quality of the results. Algorithms for highly automated spot detection were developed and many spots in 2-DE gels of *E. coli* and human plasma could be identified. Otherwise, the user can launch data mining applications and search the data for some predefined properties like modified peptide masses. As a result, hypotheses can be checked or statistical tasks can be performed. Current development focuses on automation, data mining and visualisation. We are working on the automation of criteria to exclude falsely matched proteins using the spatial correlation of the data. To gain more insight into the data, better visualisation techniques incorporating more user interaction, 3D-animation and multidimensionality will be required. It should also be possible to compare different scans and automatically retrieve information on differently expressed proteins.

In combination with a MALDI-TOF/TOF spectrometer [190] the molecular scanner approach could be used to identify proteins with peptide fragmentation data. We believe that this is a very promising approach for future development.

5. CONCLUSIONS

This chapter has dealt with the whole process of protein identification using mass spectrometry, starting with protein separation followed by mass spectra acquisition and the analysis of this data with bioinformatic tools. Examples of these techniques applied to real biological samples have been given. However, all the steps involved should be optimised. For example, various purification techniques are available and amongst them 2-D gel electrophoresis is one of the most powerful, however, only the most abundant proteins are seen. This problem can be avoided by using narrow range IPG gradients or by pre-fractionation of the sample, for example by affinity chromatography, before electrophoresis.

A direct consequence of the completion of so many genome sequencing projects is that the databases containing protein sequence information are growing exponentially. In order to have experimental data that is discriminative enough to unambiguously identify a protein, more experimental information has to be added to mass spectra data, and the available bioinformatics tools must consider this information. To enlarge the scope of mass spectrometry, data such as sequence tags obtained from MS/MS or PSD-MS measurements or information resulting from chemically modified peptides should be also taken into account.

All the processes from separation to identification should be automated. The current approach using robotics has a limited throughput because not all of its steps

can be automated or paralleled. The molecular scanner technique could circumvent these drawbacks due to its high capacity for automation and parallel processing. In combination with a MALDI-TOF-TOF MS instrument it provides a very powerful and rapid means to analyse proteins separated on a 2-DE gel.

Proteomics is not only concerned with the identification of proteins but also their control and function. This will involve considerable efforts for the foreseeable future.

6. ACKNOWLEDGEMENTS

We would like to thank Diego Chiappe, Veronique Converset, Isabelle Demalte, Jacques Deshusses, Roberto Fabbretti, Irene Fasso, Severine Frutiger-Hughes, Christine Hoogland, Ivan Ivanyi, Sylviane Jacoud, Salvo Paesano, Luisa Tonnella, Catherine Zimmermann for their technical assistance.

7. REFERENCES

1. Anderson L, Seilhamer J. *Electrophoresis* 18: 533, 1997
2. Link A, Tempel K, Hund M. *Z Naturforsch [C]* 47: 249, 1992
3. Williams KL, Hochstrasser DF. *Proteome Research: New Frontiers in Functional Genomics.* Wilkins MR, Williams KL, Appel RD, Hochstrasser DF, Eds. Berlin : Springer, 1997
4. Klose J. *Humangenetik* 26: 231, 1975
5. O'Farrell PH. *J. Biol. Chem.* 250:4007, 1975
6. Anderson NG, Anderson L. *Clin. Chem.* 28: 739, 1982
7. Taylor J, Anderson NL, Scandora AE Jr, Willard KE, Anderson NG. *Clin. Chem* 28: 861, 1982
8. Schwert GW, Takenaka Y *Biochim. Biophys. Acta* 16: 570, 1955
9. Edman P, Begg G. *Eur. J. Biochem.* 1: 80, 1967
10. Kollisch GV, Lorenz MW, Kellner R, Verhaert PD, Hoffmann KH. *Eur J Biochem* 267: 5502., 2000
11. Lehr S, Kotzka J, Herkner A, Sikmann A, Meyer HE, Krone W, Muller-Wieland D. *Biochemistry* 39: 10898, 2000
12. Ramsay SL, Steinborner ST, Waugh RJ, Dua S, Bowie JH. *Rapid Commun Mass Spectrom* 9: 11241, 1995
13. Haynes PA, Sheumack D, Greig LG, Kibby J, Redmond JW. *J. Chromatogr.* 588: 107, 1991
14. Einarsson S, Josefsson B, Lagerkvist S. *J. Chromatogr.* 282: 609, 1983
15. Balnkenship DT, Krivenek MA, Ackermann BL, Cardin AD. *Anal. Biochem.* 178: 227, 1989
16. Yan JX, Wilkins MR, Ou K, Gooley AA, Williams KL, Sanchez JC, Golaz O, Pasquali C, Hochstrasser DF. *J Chromatogr A.* 736: 291, 1996
17. Golaz O, Wilkins MR, Sanchez JC, Appel RD, Hochstrasser DF, Williams KL. *Electrophoresis* 17: 573, 1996
18. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I, Williams KL, Hochstrasser DF. *Biotechnology* 14: 61, 1996
19. Wilkins MR, Gasteiger E, Sanchez JC, Appel RD, Hochstrasser DF. *Curr Biol 6:* 1543, 1996
20. Wilkins MR, Ou K, Appel RD, Sanchez JC, Yan JX, Golaz O, Farnsworth V, Cartier P, Hochstrasser DF, Williams KL, Gooley AA. *Biochem. Biophys. Res. Commun.* 221: 609, 1996
21. Gooley AA, Ou K, Russell J, Wilkins MR, Sanchez JC, Hochstrasser DF, Williams KL. *Electrophoresis* 18: 1068, 1997
22. Bjellqvist B, Ek K, Righetti PG, Gianazza E, Görg A, Westermeier R, Postel W. *J Biochem Biophys Methods* 6: 317, 1982
23. Henzel WJ, Bililleci TN, Stults JT, Wong SC, Grilmey C, Watanabe C. *Proc. Natl. Acad. Sci. USA* 90: 5011, 1993

24. James P, Quadroni M, Carafoli E, Gonnet G. *Biochem. Biophys. Res. Commun.* 195: 58, 1993
25. Mann M, Hojrup P, Roepstorff P. *Biol. Mass Spectrum.* 22: 338, 1993
26. Pappin DJC, Hojrup P, Bleasby AJ. *Curr. Biol.* 3: 327, 1993
27. Yates JR III, Speicher S, Griffin PR, Hunkapiller T. Anal. Biochem: 214, 397, 1993
28. Karas M, Hillenkamp F. *Anal. Chem.* 60: 2299, 1988
29. Yamashita M, Fenn JB *J. Phys. Chem.* 88: 4451, 1984
30. Aleksandrov ML, Gall LN, Krasnov VN, Nikolaev VI, Pavlenko VA, Shkurov VA. *Dokl. Akad. Nauk SSSR* 277: 379, 1984
31. Rossier JS, Schwarz A, Reymond F, Ferrigno R, Bianchi F, Girault HH. *Electrophoresis* 20: 727, 1999
32. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR 3rd. *Nat Biotechnol.* 17: 676, 1999
33. Oda Y, Huang K, Cross FR, Cowburn D, Chait BT. *Proc. Natl. Acad. Sci USA* 96: 6591, 1999
34. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. *Nat Biotechnol* 17: 994, 1999
35. Kenrick KG, Margolis J. *Anal. Biochem.* 33: 1970, 204
36. Scheele GA. *J. Biol. Chem.* 250: 5375, 1975
37. Görg A, Postel W, Gunther S. *Electrophoresis* 9: 531, 1988
38. Tonella L, Sanchez J-C, Binz P-A, Appel RD, Bairoch A, Hoogland C, Hochstrasser DF. Swiss Electrophoresis Meeting: Basel (CH) 1998
39. Lämmli UK. *Nature* 277: 680, 1970
40. Rabilloud T. *Electrophoresis* 11: 785, 1990
41. Neuhoff V, Arold N, Taube D, Ehrhardt W. *Electrophoresis* 9: 255, 1988
42. Fernandez-Patron C, Castellanos-Serra L, Rodriguez P. *Biotechniques* 12: 564, 1992
43. Patton WF. *Electrophoresis* 21: 1123, 2000
44. Johnston RF, Pickett SC, Barker DL. *Electrophoresis* 11: 355, 1990
45. Appel RD, Hochstrasser DF, Funk M, Vargas JR, Pellegrini C, Muller AF, Scherrer JR. *Electrophoresis* 12: 722, 1991
46. Walsh BJ, Molloy MP, Williams KL. *Electrophoresis* 19: 1883, 1998
47. Traini M, Gooley AA, Ou K, Wilkins MR, Tonella L, Sanchez J-C, Hochstrasser DF, Williams KL. *Electrophoresis* 19: 1941, 1998
48. Houthaeve T, Gausepohl H, Mann M, Ashman K. *FEBS Lett.* 376: 91, 1995
49. Houthaeve T, Gausepohl H, Ashman K, Nillson T, Mann M. *J Protein Chem* 16: 343, 1997
50. Ashman K, Houthaeve T, Clayton J, Wilm M Podtelejnikov A, Jensen ON. *Lett. Pept. Sci.* 4: 244, 1997
51. Gras R, Müller M, Gasteiger E, Gay S, Binz P-A, Bienvenut W, Hoogland C, Sanchez J-C, Bairoch A, Hochstrasser DF, Appel R. *Electrophoresis* 20: 3535, 1999
52. Breen EJ, Hopwood FG, Williams KL, Wilkins MR. *Electrophoresis* 21: 2243, 2000
53. Tanaka K, Waki h, Ido Y, Akita S, Yoshida Y, Yoshida T *Rapid Commun. Mass Spectrom* 2: 151, 1988
54. Li G, Waltham M, Anderson NL, Unsworth E, Treston A, Weinstein JN. *Electrophoresis* 18: 391, 1997
55. Miliotis T, Kjellstrom S, Nilsson J, Laurell T, Edholm LE, Marko-Varga G. *J. Mass Spectrom.* 35:369, 2000
56. Nakanishi T, Okamoto N, Tanaka K, Shimizu A. *Biol Mass Spectrom* 23: 230, 1994
57. Okamoto M, Takahashi K, Doi T, Takimoto Y. *Anal Chem* 69: 2919, 1997
58. Stemmler EA, Hettich RL, Hurst GB, Buchanan MV *Rapid Commun Mass Spectrom* 7: 828,1993
59. Stemmler EA, Buchanan MV, Hurst GB, Hettich RL. *Anal Chem* 67: 2924, 1995
60. Gusev AI, Wilkinson WR, Proctor A, Hercules DM. *Anal. Chem.* 67: 1034, 1995
61. Karas M, Bahr U, Strupat K, Hillenkamp F. *Anal. Chem.* 67: 675, 1995
62. Taranenko NI, Tang K, Allman KL, ChangLY, Chen CH *Rapid Comm. Mass. Spectrom* 8: 1001, 1994
63. Zhu YF, Chung CN, Taranenko NI, Allman SL, Martin SA, Haff L, *Chen CH. Rapid Comm. Mass Spectrom.* 10: 383, 1996
64. Ayorinde FO, Hambright P, Porter TN, Keith QL Jr. *Rapid Commun Mass Spectrom* 13: 2474, 1999
65. Wei J, Buriak JM, Siuzdak G *Nature 399*: 243,1999
66. Gonnet GH. "A tutorial introduction to Computational Biochemistry using Darwin", Technical Report, E.T.H Zurich, Switzerland, November, 1992

67. Pappin DJC, Hojrup P, Bleasby AJ. *Curr. Biol.,* 3: 327, 1993.
68. Clauser KR, Hall SC, Smith DM, Webb JW, Andrews LE, Tran HM, Epstein LB, Burlingame AL. *Proc. Natl. Acd. Sci. USA* 92: 5072, 1995
69. Ingendoh, 1994
70. Mamyrin BA, Karatajev VJ, Shmikk DV, Zagulin V. *JETP* 37: 45, 1973
71. Vestal ML, Juhasz P, Martin SA *Rapid Commun.Mass Spectrom* 9: 1044, 1995
72. Whittal RM, LI L *Anal. Chem.* 67: 1950, 1995
73. Brown RS, Lennon JJ. *Anal. Chem.* 67: 1998, 1995
74. Jensen ON, Podtelejnikov A, Mann M. *Rapid Commun. Mass Spectrom* 10: 1371, 1996
75. Wiley WC, McLaren IH. Rev. *Sci. Instrum.* 26: 1150, 1953
76. Takach EJ, Hines WM, Patterson DH, Juhasz P, Falick AM, Vestal ML, Martin SA. *J. Prot. Chem.* 16: 363, 1997
77. Chen C, Walkes AK, Wu Y, Timmons RB, Kinsel GR. *J. Mass Spectrom.* 34: 1205, 1999
78. Cohen SL, Chait BT. *Anal. Chem.* 68: 31, 1996
79. Figueroa ID, Torres O, Russell DH. *Anal. Chem.* 70: 4527, 70
80. Kratzer R, Eckerskorn C, Karas M, Lottspeich F. *Electrophoresis* 19: 1910, 1998
81. Beavis R, Bridson JN. *J. Phys. D: Appl. Phys.* 26: 442, 1993
82. Wenschuh H, Halada P, Lamer S, Jungblut P, Krause E *Rapid Commun. Mass Spectrom.* 12: 115, 1998
83. Amado FML, Domingues P, Santana-Marques MG, Ferrer-Correia AJ, Tomer KB. *Rapid Commun. Mass Spectrom.* 11: 1347, 1997
84. Patterson SD, Thomas D, Bradshaw RA. *Electrophoresis* 17:877, 1996
85. Breaux GA, Green-Church KB, France A, Limbach PA. *Anal. Chem.* 72: 1169, 2000
86. Krause E, Wenschuh H, Jungblut PR. *Anal. Chem.* 71: 4160, 1999
87. Keil B. Specificity of proteolysis. Heidelberg, New York: Springer-Verlag, 1992,
88. Gobom J, Kraeuter KO, Persson R, Steen H, Roepstorff P, Ekman R. *Anal. Chem. 72*: 3320, 2000
89. Hensel RR, King RC, Owens KG. *Rapid Commun. Mass Spectrom.* 11: 1785, 1997
90. Kaufmann R, Spengler B, Lützenkirchen F. *Rapid Commun. Mass Spectrom.* 7: 902, 1993
91. Kaufmann R, Kirsch D, Spengler B. *Int. J. Mass Spectrom. Ion Processes* 131: 355, 1994
92. Roepstorff P, Fohlmann J. *Biomed. Mass Spectrom.* 11: 601, 1984
93. Johnson RS, Martin SA, Biemann K. *Int. J. Mass Spectrom. Ion Processes* 86: 137, 1988
94. Spengler B *J. Mass Spectrom.* 32: 1019, 1997
95. Spengler B, Lutzenkirchen F, Kaufmann R *Organ. Mass Spectrom.* 28: 1482, 1993
96. Liao PC, Huang ZH, Allison J. *J. Am. Soc. Mass Spectrom* 8: 501, 1997
97. Spengler B, Luetzenkirchen F, Metzger S, Chaurand P, Kaufmann R, Jeffery W, Bartlet-Jones M, Pappin DJC. *Int. J. Mass Spectrom. Ion Processes* 169/170: 127, 1997
98. Bauer MD, Sun Y, Keough T, Lacey MP. *Rapid Commun. Mass Spectrom.* 14: 924, 2000
99. Keough T, Youngquist RS, Lacey MP. *Proc. Natl. Acad. Sci. USA* 96: 7131, 1999
100. Keough T, Lacey MP, Fieno MA, Grant RA, Sun Y, Bauer MD, Begley KB. *Electrophoresis* 21: 2252, 2000
101. Gevaert K, Demol H, Puype M, Broekaert D, De Boeck S, Houthaeve T, Vandekerckhove J. *Electrophoresis* 18: 2950, 1997
102. Gevaert K, De Mol H, Puype M, Houthaeye T, De Boeck S, Vandekerckhove J. *J. Protein Chem.* 17: 560, 1998
103. Gevaert K, Demol H, Sklyarova T, Vandekerckhove J, Houthaeve T. *Electrophoresis* 19: 909, 1998
104. Gevaert K, Vandekerckhove J. *Electrophoresis* 21: 1145, 2000
105. Cornish TJ, Cotter RJ. *Rapid. Commun. Mass Spectrom.* 7: 1037,1993
106. Medzihradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML, Burlingame AL. *Anal. Chem.* 72: 552, 2000
107. Vestal ML, Campbell JM, Hayden K, Juhasz P *Performance Evaluation of Improved MALDI TOF-TOF MS System.* Proceedings of 48th: Long Beach (CA), ASMS 2000
108. Krutchinsky AN, Zhang W, Chait BT. *J. Am. Soc. Mass Spectrom.* 11: 493, 2000
109. Shevchenko A, Loboda A, Shevchenko A, Ens W, Standing KG. *Anal. Chem.* 72: 2132, 2000
110. Verentchikov AN, Hayden K, Vestal ML. *Tandem TOF-Orthogonal TOF Mass Spectrometer with MALDI Ion Source.* Proceedings of 48th: Long Beach (CA), ASMS 2000
111. Smith RD, Loo JA, Edmonds CG, Barinaga CJ, Udseth HR. *Anal. Chem.* 62: 882, 1990

112. Aleksandrov ML, Gall LN, Krasnov VN, Nikolae VI, Pavlenko VA, Shkurov VA, Baram GI, Gracher MA, Knorre VD, Kusner VS. *Bioorg. Khim.* 10: 710, 1984
113. Kebarle P, Peschke M. *Anal. Chim. Acta* 406: 11, 2000
114. Guilhaus M, Selby D, Mlynski V. *Mass Spectrom. Rev.* 19: 65, 2000
115. Wilm M, Mann M *Anal. Chem.* 68: 1, 1996
116. Emmett MR, Caprioli RM. *J. Am. Soc. Mass Spectrom.* 5: 605, 1994
117. Ducret A, van Oostveen I, Eng JK, Yates III JR, Aebersold R. *Protein Sci.* 7: 706, 1998
118. Lahm H-M, Langen H. *Electrophoresis* 21: 2105,2000
119. Lopez MF *Electrophoresis* 21: 1082, 2000
120. Bienvenut *et al.* In preparation
121. Fraenkem-Conrat H, Olcott HS. *J. Biol. Chem.* 161: 259, 1945
122. Hunt DF, Yates JR, Shabanowitz J, Winston S, Hauer CR. *Proc. Natl. Acad. Sci. USA* 83: 6233, 1986
123. Wilcox PE. *Meth. in Enzym.* 11: 605, 1967
124. Bartlet-Jones M, Hansen H, Pappin DJC. *Rap. Comm. Mass Spectrom.* 8: 737,1994
125. Acharya AS, Maanjula BN, Murthy GS, Vithayathil PJ. *Int. J. Protein Res.* 9: 1977, 213
126. Falick AM, Maltby DA. *Anal Biochem* 182: 1989, 165
127. Nutkins JC, Williams DH. *Eur. J. Biochem* 181: 1989, 97
128. Jones DD, Stott KM, Howard MJ, Perham RN. *Biochemistry* 39: 8448, 2000
129. Wang F, Tang X *Biochemistry* 35: 4069, 1996
130. Katta V, Chait BT. *J. Am. Chem. Soc.* 115: 6317, 1993
131. Villanueva J, Canals F, Villegas V, Querol E, Aviles FX. *FEBS Lett* 472: 27, 2000
132. Zhang W Chait BT. *Anal. Chem* 72: 2482, 2000
133. Kraus M, Janek K, Bienert M, Krause E. *Rapid Commun Mass Spectrom* 14: 1094, 2000
134. Buijs J, Costa Vera C, Ayala E, Steensma E, Hakansson P, Oscarsson S. *Anal Chem* 71: 3219, 1999
135. Chaurand P, Luetzenkirchen F, Spengler B. *J. Am. Soc. Mass Spectrom.* 10: 91, 1999
136. Thiede B, Siejak F, Dimmler C, Jungblut PR, Rudel T. *Electrophoresis* 21: 2713, 2000
137. Joubert-Caron R, Le Caer JP, Montandon F, Poirier F, Pontet M, Imam N, Feuillard J, Bladier D, Rossier J, Caron M. *Electrophoresis* 21: 2566, 2000
138. Hochstrasser DF. *Clin. Chem. Lab. Med.* 36: 825, 1998
139. Binz P-A, Müller M, Walther D, Bienvenut WV, Gras R, Hoogland C, Bouchet G, Gasteiger E, Fabbretti R, Gay S, Palagi P, Wilkins M, Rouge V, Tonella L, Paesano S, Rosselat G, Karmime A, Bairoch A, Sanchez J-C, Appel RD, Hochstrasser DF. *Analytical Chemistry* 71: 4981, 1999
140. Hochstrasser DF, Appel RD, Vargas R, Perrier R, Vurlod JF, Ravier F, Pasquali C, Funk M, Pellegrini C, Muller AF, Scherrer MD. *MD Comput.* 8: 85, 1991
141. Bienvenut WV, Sanchez J-C, Karmime A, Rouge V, Rose K, Binz P-A, Hochstrasser DF. *Analytical Chemistry* 71: 4800, 1999
142. Fabbretti R, Binz PA, Bienvenut W, Gasteiger E, Bairoch A, Wilkins MR, Sanchez JC, Walther D, Hochstrasser DF, Appel RD. 3rd Siena 2D Electrophoresis Meeting: Sienna (IT), 1998
143. Jungblut P, Eckerskorn C, Lottspeich F, Klose J. *Electrophoresis* 11: 581, 1990
144. Mozdzanowski J, Speicher DW. *Anal Biochem* 207: 11, 1992
145. Reim DF, Speicher DW. *Anal Biochem* 207: 19, 1992
146. Bolt MW, Mahoney PA. *Anal Biochem* 247: 185, 1997
147. Neumann H, Mullner S. *Electrophoresis* 19: 752, 1998
148. Appel RD, Palagi PM, Walther D, Vargas JR, Sanchez JC, Ravier F, Pasquali C, Hochstrasser DF. *Electrophoresis.* 18: 2724, 1997
149. Salih B, Zenobi R. *Anal. Chem.* 70: 1536, 1998
150. Tal M, Silberstein A, Nusser E. *J. Biol. Chem.* 260: 9976, 1985
151. Dottavio-Martin D, Ravel JM. *Anal. Biochem.* 87: 562, 1978
152. Binz PA, Wilkins M, Gasteiger E, Bairoch A, Appel R and Hochstrasser DF. In Microcharacterization of Proteins. Wiley-VHC. Internet resources for protein identification and characterization, 277, 1999
153. Bairoch A, Apweiler R. *Nucleic Acids Res.* 28:45-48, 2000
154. Pappin DCJ, Hojrup P, Bleasby AJ. *Curr. Biol.* 3: 327, 1993
155. Perkins DN, Pappin DJC, Creasy DM, Cotrell JS. *Electrophoresis* 20: 3551,1999
156. Zhang W and Chait BT. *Analytical Chemistry,* 72: 2482, 2000
157. Eng JK, McCormack AL, Yates JR. *J. Am. Soc. Mass Spec.* 5:976, 1994

158. Mann M, Wilm M. *Anal. Chem.* 66: 4390, 1994
159. Fenyo D, Qin J, Chait BT. *Electrophoresis* 19: 998, 1998
160. McCormack AL, Schieltz DM, Goode B, Yang S, Barnes G, Drubin D, Yates JR 3rd. *Anal. Chem.* 69: 767, 1997
161. Yates JR 3d, Eng JK, McCormack AL, Schieltz D. *Anal. Chem.* 67: 1426, 1995
162. Sakurai T, Matsuo T, Matsuda H, Katakuse I. *Biomed. Mass Spectrom,* 11:396, 1984
163. Johnson RJ, Biemann K. *Biomed. Environ. Mass Spectrom.* 18:945, 1989
164. Yates JR, Griffin PR, Hood LE and Zhou JX. *Techniques in Protein Chemistry II.* Villafranca JJ Ed., San Diego: Academic Press, 477, 1991
165. Taylor JA, Johnson R. *Rapid Commun. Mass Spectrom.* 11:1067, 1997
166. Fernandez-de-Cossio J, Gonzales J, Besada V. *Comput Appl Biosci.* 11:427, 1995
167. Fernandez-de-Cossio J, Gonzales J, Betancourt L, Besada V, Padron G, Shimonishi Y, Takao T. *Rapid Comm. Mass. Spectrom* 12: 1867, 1998
168. Dancik V, Addona TA, Clauser KR, Vath JE, Pevtner PA. *J. Comput. Bio.* 6:327, 1999
169. Zhang Z, McElvain JS. *Anal. Chem.* 72:2337, 2000
170. Stranz DD, Martin LB. 3d. *J. Biomol. Techniques*, 9: 19, 1998
171. Scarberry RE, Zhang Z, Knapp DR. *J. Am. Soc, Mass Spectrom*, 6:947, 1995
172. Wilkins MR, Gasteiger E, Gooley AA, Herbert BR, Molloy MP, Binz PA, Ou K, Sanchez JC, Bairoch A, Williams KL, Hochstrasser DF. *J. Mol Biol.* 289: 645, 1999
173. Hofmann K, Bucher P, Falquet L, Bairoch A. *Nucl. Acids Res.* 27:215, 1999
174. Cooper CA, Gasteiger E, Packer N. *Proteomics* 1: 2001, in press
175. Avery G, McGee C, Falk S. *Anal. Chem.* 1:57A, 2000
176. Gates PJ, Kearney GC, Jones R, Leadlay PF, Staunton J. *Rapid Commun. Mass Spectrom.* 13: 242, 1999
177. Sullards MC, Reiter JA. *J Am Soc Mass Spectrom.* 11: 40, 2000
178. Thiede B, Lamer S, Mattow J, Siejak F, Dimmler C, Rudel T, Jungblut PR. *Rapid Commun. Mass Spectrom.* 14: 496, 2000
179. Vestal M, Juhasz P. *J. Am. Soc. Mass Spectrom.* 9: 892, 1998
180. Karas M, Glückmann M, Schäfer J. *J. Mass Spectrom.* 35: 1, 2000
181. Zenobi R, Knochenmuss R. *Mass Spectrom. Rev.* 17: 337, 1998
182. Gay S, Binz PA, Hochstrasser DF, Appel RD. *Electrophoresis* 20: 3527, 1999
183. Hoogland C, Sanchez JC, Tonella L, Binz PA, Bairoch A, Hochstrasser DF, Appel RD. *Nucleic Acids Res.* 28: 286, 2000
184. Cavalcoli JD, VanBogelen RA, Andrews PC, Moldower B. *Electrophoresis* 18: 2703, 1997
185. Eriksson J, Chait BT, Fenyö D. *Anal. Chem.* 72: 999, 2000
186. Toffoli T, Margolus N. *Cellular Automata Machines.* Cambridge (MA): MIT Press, 1987
187. Keller BO, Li L. *J. Am. Soc. Mass Spectrom.* 11: 88, 2000
188. Müller M *et al.* In preparation
189. Saporta G. Probabilités, Analyse des Données et Statistique. Paris : Editions Technip, 1990
190. Medzuhradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML, Burlingame AL. *Anal. Chem.* 72: 552, 2000
191. Saporta G. Probabilités, Analyse des Données et Statistique. Paris : Editions Technip, 1990
192. Medzuhradszky KF, Campbell JM, Baldwin MA, Falick AM, Juhasz P, Vestal ML, Burlingame AL. *Anal. Chem.* 72:552, 2000
193. Bienvenut WV, Déon C, Sanchez JC, Hochstrasser DF Electrophoresis submitted, 2000

INDEX